

Weakly Supervised Learning of Objects and Attributes



ZHIYUAN SHI

School of Electronic Engineering and Computer Science

Queen Mary, University of London

A thesis submitted for the degree of

Doctor of Philosophy

February 2016

Abstract

This thesis presents weakly supervised learning approaches to directly exploit image-level tags (e.g. objects, attributes) for comprehensive image understanding, including tasks such as object localisation, image description, image retrieval, semantic segmentation, person re-identification and person search, etc. Unlike the conventional approaches which tackle weakly supervised problem by learning a discriminative model, a generative Bayesian framework is proposed which provides better mechanisms to resolve the ambiguity problem. The proposed model significantly differentiates from the existing approaches in that: (1) All foreground object classes are modelled jointly in a single generative model that encodes multiple objects co-existence so that “explaining away” inference can resolve ambiguity and lead to better learning. (2) Image backgrounds are shared across classes to better learn varying surroundings and “push out” objects of interest. (3) the Bayesian formulation enables the exploitation of various types of prior knowledge to compensate for the limited supervision offered by weakly labelled data, as well as Bayesian domain adaptation for transfer learning.

Detecting objects is the first and critical component in image understanding paradigm. Unlike conventional fully supervised object detection approaches, the proposed model aims to train an object detector from weakly labelled data. A novel framework based on Bayesian latent topic model is proposed to address the problem of localisation of objects as bounding boxes in images and videos with image level object labels. The inferred object location can be then used as the annotation to train a classic object detector with conventional approaches.

However, objects cannot tell the whole story in an image. Beyond detecting objects, a general visual model should be able to describe objects

and segment them at a pixel level. Another limitation of the initial model is that it still requires an additional object detector. To remedy the above two drawbacks, a novel weakly supervised non-parametric Bayesian model is presented to model objects, attributes and their associations automatically from weakly labelled images. Once learned, given a new image, the proposed model can describe the image with the combination of objects and attributes, as well as their locations and segmentation.

Finally, this thesis further tackles the weakly supervised learning problem from a transfer learning perspective, by considering the fact that there are always some fully labelled or weakly labelled data available in a related domain while only insufficient labelled data exist for training in the target domain. A powerful semantic description is transferred from the existing fashion photography datasets to surveillance data to solve the person re-identification problem.

Dedication

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged.

Some parts of the work have previously been published as:

- Chapter 3
 - Zhiyuan Shi, Timothy M. Hospedales and Tao Xiang, “Bayesian Joint Topic Modelling for Weakly Supervised Object Localisation”, in International Conference on Computer Vision (ICCV), 2013
 - Zhiyuan Shi, Timothy M. Hospedales and Tao Xiang, “Bayesian Joint Modelling for Object Localisation in Weakly Labelled Images”, in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015
- Chapter 4
 - Zhiyuan Shi, Yongxin Yang, Timothy M. Hospedales and Tao Xiang, “Weakly Supervised Learning of Objects, Attributes and their Associations”, European Conference on Computer Vision (ECCV), 2014
 - Zhiyuan Shi, Yongxin Yang, Timothy M. Hospedales and Tao Xiang, “Weakly Supervised Image Annotation and Segmentation with Objects and Attributes”, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015

- Chapter 5
 - Zhiyuan Shi, Timothy M. Hospedales and Tao Xiang, “Transferring a Semantic Representation for Person Re-Identification and Search”, in International Conference on Computer Vision and Pattern Recognition (CVPR), 2015

Acknowledgements

I consider myself extremely lucky to have had the opportunity to work with my supervisor, Dr. Tao Xiang. I would like to thank my supervisor for his perpetual patience, encouragement and guidance. I will never forget his kind help in every bit of the PhD process. Besides, I am deeply grateful for invaluable advice and consistent support from my co-supervisor Professor Shaogang Gong throughout my four year PhD life. I would like to thank Dr. Lourdes Agapito for being my internal examiner throughout my PhD project.

I would also like to thank Dr. Timothy Hospedales. I greatly appreciate his tremendous help on my research works. He could be considered to my third supervisor. Our collaborative efforts have led to several successes. He has continuously provided stimulating conversations which has led to an ever growing list of new ideas to try.

My warm appreciation goes to various students and associates at Vision Group for their friendship and support, in particular Parthipan Siva, Chris Russell, Yi-Zhe Song, Miles Hansard, Chen Change Loy (Cavan), Ke Chen (Cory), Zhenyong Fu (Ian), Yongxin Yang, Xiatian Zhu (Eddy), Yanwei Fu, Ryan Layne, Yi Li, Xun Xu (Alex), Hanxiao Wang, Ioannis Alexiou, Li Zhang, Elyor Kodirov, Kunkun Pang, Yaowei Wang, Xiangyu Kong, Shuxin Quyang, Jingya Wang, Qian Yu and Rui Yu.

Last but not least, I am eternally grateful to my family for their enduring love, support and understanding.

Contents

Contents	vi
List of Figures	x
List of Tables	xiii
List of Abbreviations	xv
List of Symbols	xviii
1 Introduction	1
1.1 Describing Images with Objects and Attributes	2
1.1.1 Object Recognition and Localisation	2
1.1.2 Describing Objects with Attributes	4
1.1.3 Applications	6
1.2 Challenges and Motivations	9
1.2.1 General Challenges for Object and Attribute Recognition . . .	9
1.2.2 Challenges Faced by Fully Supervised Learning based Ap- proaches	10
1.2.3 Challenges Faced by Weakly Supervised Learning based Ap- proaches	10
1.2.4 Motivations	11
1.3 The Proposed Approach	13
1.3.1 A Joint Bayesian Approach for Weakly Supervised Learning .	13
1.3.2 Weakly Supervised Localisation of Objects	14
1.3.3 Weakly Supervised Learning of Objects and Attributes	16

1.3.4	Transferring Learning of Objects and Attributes	19
1.4	Contributions	21
1.5	Outline	22
2	Literature Review	23
2.1	Object Detection	23
2.1.1	Fully Supervised Object Detection	23
2.1.2	Weakly Supervised Object Detection	26
2.2	Object Segmentation	27
2.2.1	Fully Supervised Segmentation	27
2.2.2	Weakly Supervised Segmentation	29
2.3	Attribute Learning	30
2.3.1	Attribute Predication and Localisation	30
2.3.2	Attribute as Representation for Vision Tasks	31
2.4	Joint Learning of Object and Attributes.	33
2.5	Related Work in Weakly Supervised Learning	34
2.5.1	Discriminative Models	34
2.5.2	Probabilistic Generative Models	35
2.5.3	Exploiting Prior Knowledge	36
2.5.4	Cross Domain Learning	37
2.5.5	Exploiting Unlabelled Data	38
2.5.6	Feature Fusion	38
2.6	Summary	39
3	Bayesian Joint Modelling for Object Localisation	42
3.1	Overview	42
3.2	Joint Topic Model for Objects and Background	43
3.3	Relationship with Latent Dirichlet Allocation	46
3.4	Model Learning	47
3.5	Object Localisation	50
3.6	Bayesian Priors	51
3.7	Learning from Additional Data	53
3.7.1	Bayesian Domain Adaptation	53

3.7.2	Semi-supervised Learning	54
3.8	Experiments	54
3.8.1	Datasets, Features and Settings	54
3.8.2	Comparison with State-of-the-art	56
3.8.2.1	Results on VOC Dataset	56
3.8.2.2	Results on ImageNet Dataset	63
3.8.2.3	Results on YouTube-object Dataset	64
3.8.3	Bayesian Domain Adaptation	66
3.8.4	Semi-supervised Learning	68
3.8.5	Computational Cost	69
3.9	Summary	70
4	Weakly Supervised Learning of Objects, Attributes and their Associations	72
4.1	Overview	72
4.2	Weakly Supervised Stacked Indian Buffet Process	74
4.2.1	Image Representation	74
4.2.2	Model Formulation	75
4.2.2.1	WS-SIBP	76
4.2.2.2	WS-MRF-SIBP	78
4.2.2.3	Comparison with Joint Topic Model	79
4.2.3	Model Learning	80
4.2.4	Inference for Test Data	82
4.2.5	Applications of the Model	82
4.3	Experiments	83
4.3.1	Image Annotation and Query	84
4.3.1.1	Datasets and Settings	84
4.3.1.2	Image Annotation	86
4.3.1.3	Object-attribute Query	90
4.3.2	Semantic Segmentation	91
4.3.2.1	Datasets and Settings	91
4.3.2.2	Results on aPascal	92
4.3.2.3	Results on LabelMe	95
4.3.3	Further Evaluations	98

4.4	Summary	99
5	Transferring a Semantic Representation for Person Re-identification and Search	101
5.1	Overview	101
5.2	Semantic Representation Learning	103
5.2.1	Model Formulation	104
5.2.2	Model Learning from the Auxiliary Set	106
5.2.3	Model Adaptation to the Target Set	107
5.3	Semantic Representation Applications	107
5.3.1	Person Re-identification	107
5.3.2	Person Search	108
5.4	Experiments	108
5.4.1	Datasets and Settings	108
5.4.2	Person Re-identification	111
5.4.3	Person Search	114
5.5	Summary	116
6	Conclusion and Future Work	117
Appendix A	Detailed Results for Object Localisation	121
A.1	Further Evaluations	121
A.2	Per-Class Object Localisation Results	124
A.3	Per-Class Object Detection Results	126
Appendix B	Detailed Results for Re-identification and Search	129
B.1	Details on Supervision	129
B.2	Re-identification Performance Measured by CMC	131
B.3	Per-query Person Search Results	131
	References	135

List of Figures

1.1	An example of real world images can be easily described by a young child. [1].	2
1.2	Visual understanding in real world scenes. Given a real-world image, an ideal intelligent vision system can recognise the objects within an image, such as a tree, a person. It can further describe these objects in detail by their characteristics. For example, a green tree, a group of people who are talking with each other. Moreover, it is able to identify the landmark (i.e. This is Queen’s building of Queen Mary, University of London).	3
1.3	An example of object detection results from trained detector [2]. . . .	4
1.4	An example of object segmentation results from trained model [3]. . .	5
1.5	Examples of different kinds of attributes [4].	6
1.6	An example of generating detailed description of images [5].	7
1.7	An example of image search given a multi-attribute query [6].	8
1.8	An illustration of person re-identification	8
1.9	An illustration of person search	9
1.10	Weakly supervised object localisation	15
1.11	Weakly supervised learning of objects and attributes	17
1.12	Transferring a Semantic Representation	20
3.1	Different types of objects often co-exist in a single image. The proposed joint learning approach differs from previous approaches which localise each object class independently.	44
3.2	Graphical model for the proposed WSOL joint topic model. Shaded nodes are observed.	45

LIST OF FIGURES

3.3	Graphical model for the classic latent dirichlet allocation. Shaded nodes are observed.	47
3.4	(a) A hierarchical structure of the 20 PASCAL VOC classes using WordNet. (b) The class similarity matrix.	59
3.5	Top row in each subfigure: examples of object localisation using our-sampling and our-Gaussian. Bottom row: illustration of what is learned by the object (foreground) topics via heat map (brighter means object is more likely). The first four rows show some examples of PASCAL VOC and last two rows are selected from ImageNet.	60
3.6	Illustration of the learned background topics.	62
3.7	Examples of video object localisation	65
3.8	Domain adaptation provides more benefit with fewer target domain samples. Specifically, the presented model ($V \rightarrow Y$ and $Y \rightarrow V$) provides a bigger margin of benefit given less target domain data.	67
3.9	Unlabelled data improves foreground heat maps.	70
4.1	Comparing the proposed Weakly Supervised (WS) approach to object-attribute association learning to the conventional strongly supervised approach.	74
4.2	The probabilistic graphical model representing the proposed WS-MRF-SIBP. Shaded nodes are observed.	75
4.3	Strong bounding-box-level annotation and weak image-level annotations for aPascal are used for learning strongly supervised models and WS models respectively.	84
4.4	43 subordinate classes of dog are converted into a single entry-level class ‘dog’.	84
4.5	Qualitative results on free annotation. False positives are shown in red. If the object prediction is wrong, the corresponding attribute box is shaded.	86
4.6	Illustrating the inferred patch-annotation. Object and attributes are coloured, and multi-label annotation blends colours. The bottom two groups each have two rows corresponding to the two most confident objects detected.	89

LIST OF FIGURES

4.7	Object-attribute query results as precision-average recall curve.	90
4.8	Object-attribute query: qualitative comparison	91
4.9	Qualitative illustration of (attribute-enhanced) semantic segmentation results on aPascal.	94
4.10	Qualitative comparison of the proposed semantic segmentation versus alternatives on the LabelMe dataset.	97
4.11	Object-attribute query results as precision-average recall curve.	100
5.1	Transferring knowledge from fashion data to surveillance	102
5.2	Illustration of surveillance person search procedure	103
5.3	The transfer learning framework.	104
5.4	Visualisation of the proposed model output. Each patch is colour-coded to show the inferred dominant attribute of two types.	111
5.5	Person search qualitative results. The top ranked images for each query are shown. Red boxes are false detections.	114
5.6	Person search: comparison with state-of-the-art.	114
A.1	Comparing different feature fusion methods.	122
B.1	CMC comparison of unsupervised learning based approaches.	130
B.2	CMC comparison of supervised learning based approaches.	130
B.3	Person search performance on each object-attribute query	133

List of Tables

3.1	Comparison with state-of-the-art competitors on the three variations of the PASCAL VOC 2007 dataset. * Requires aspect ratio to be set manually. + Require 10 out of the 20 classes fully annotated with bounding-boxes and used as auxiliary data.	57
3.2	Performance of strong detectors trained using annotations obtained by different WSOL methods	62
3.3	Initial annotation accuracy on ImageNet dataset	63
3.4	Performance comparison on YouTube-object	65
3.5	Cross-domain transfer learning results. The proposed transfer learning strategy ($V \rightarrow Y$ and $Y \rightarrow V$) shows substantial improvements over the standard combinations ($A \rightarrow Y$ and $A \rightarrow V$) and the baselines (Y and V).	67
3.6	Localisation performance of semi-supervised learning. Unlabelled data helps to learn a better object model.	69
4.1	Free annotation performance evaluated on t attributes per object.	87
4.2	Results on annotation given object names (GN) or locations (GL). SS stands for Strongly Supervised.	89
4.3	Quantitative semantic segmentation comparison versus state-of-the-art on the aPascal dataset.	93
4.4	Quantitative comparison of semantic segmentation performance on the LabelMe dataset.	95

LIST OF TABLES

4.5	Evaluation of individual components of the proposed model. This table reports AP@2 for free annotation on aPascal and ImageNet dataset. For segmentation, per-pixel results are reported for LabelMe and IOU for aPascal. For segmentation result on aPascal 8 attribute annotations are used.	98
5.1	Matching accuracy @ rank r (%): unsupervised learning approaches. ‘-’ indicates no result was reported and no code is available for implementation. The best results for single-cue and fused-cue methods are highlighted in bold separately.	109
5.2	Matching accuracy @ rank r (%): supervised learning approaches on re-identification.	112
5.3	Effects of auxiliary data source and annotation.	112
5.4	Contribution of each model component	113
5.5	Comparing different transfer learning approaches	113
A.1	Evaluation of individual components of the proposed model and comparison with alternative joint learning approaches.	123
A.2	Per-class localisation accuracy for the <i>VOC07-6×2</i> dataset	125
A.3	Per-class localisation accuracy for the <i>VOC07-14</i> dataset	125
A.4	Per-class localisation accuracy for the <i>VOC07-20</i> dataset	126
A.5	Per-class average precision for object detection on <i>VOC07-6×2</i> dataset	127
A.6	Per-class average precision for object detection on <i>VOC07-20</i> dataset .	128
B.1	Different types of supervision used by the proposed models depend on how different attributes of each auxiliary dataset are annotated.	129

List of Abbreviations

AP Average Precision. [87](#)

BGP Background Patch. [104](#), [123](#), [130](#)

BoW Bag-of-Words. [16](#), [25](#), [43](#), [86](#)

CMC Cumulative Match Characteristic. [111](#)

CNN Convolutional Neural Network. [25](#)

CRF Conditional Random Field. [26](#), [28–30](#), [34](#)

DPM Deformable Part Model. [24](#)

FGP Foreground Patch. [110](#), [130](#)

FS Fully Supervised. [23](#), [26](#), [29](#), [30](#), [39](#), [40](#), [43](#), [56](#), [61](#), [85](#), [95](#), [97](#), [126](#)

FSL Fully Supervised Learning. [10](#), [11](#), [25](#), [39](#)

IBP Indian Buffet Process. [18](#), [21](#), [22](#), [29](#), [35](#), [36](#), [40](#), [102](#)

IL Independent Learning. [12](#), [122](#)

IOU Intersection-over-union. [92](#)

LBP Local Binary Pattern. [55](#)

LDA Latent Dirichlet Allocation. [36](#), [46](#), [85](#)

- LTM** Latent Topic Model. [16](#), [22](#), [35](#), [43](#)
- MAP** Mean Average Precision. [89](#)
- MAR** Mean Average Recall. [90](#)
- MIL** Multi-instance Learning. [15](#), [26](#), [27](#), [34](#)
- MIML** Multi-instance Multi-label. [26](#), [36](#), [85](#), [123](#)
- MKL** Multiple Kernel Learning. [37](#), [39](#)
- MRF** Markov Random Field. [18](#), [36](#)
- NM** Negative Mining. [56](#)
- NMS** Non-maximum Suppression. [50](#)
- OS** Objective Saliency. [56](#)
- PR** Precision Recall. [89](#), [114](#), [131](#)
- PTM** Probabilistic Topic Model. [35](#)
- Re-ID** Re-identification. [8](#), [19–21](#), [31](#), [102](#), [104](#), [107](#), [110](#), [111](#), [113](#), [116](#), [118](#)
- SSL** Semi-supervised Learning. [16](#), [35](#), [38](#), [70](#)
- SVM** Support Vector Machine. [27](#)
- VMP** Variational Message Passing. [48](#)
- WS** Weakly Supervised. [xi](#), [12](#), [13](#), [17](#), [22](#), [23](#), [27](#), [29](#), [34](#), [36](#), [39](#), [40](#), [49](#), [56](#), [60](#), [61](#), [64](#), [66](#), [72–74](#), [81](#), [83–93](#), [95](#), [97](#), [101](#), [118](#), [124](#), [126](#), [127](#), [130](#), [131](#)
- WS-MRF-SIBP** Weakly Supervised Markov Random Field Stacked Indian Buffet Process. [17](#), [18](#), [35](#), [79](#), [80](#), [118](#)
- WS-SIBP** Weakly-supervised Stacked Indian Buffet Process. [22](#), [73](#), [76](#)

List of Abbreviations

WSL Weakly Supervised Learning. 11–13, 15, 19, 23, 25, 26, 34, 39, 40, 76, 95, 99, 101, 118, 120

WSOL Weakly Supervised Object Localisation. 10, 12, 14, 21, 22, 26, 35–38, 42, 43, 46, 54, 60, 70, 85, 117, 118, 124

YTO YouTube-Object. 14, 54, 64, 66

List of Symbols

- C Number of classes. 43, 44
- F Maximum number of different feature types. 44, 45
- H Latent hidden variables. 46
- J Set of training images. 44, 45, 76, 104
- K Number of topics or latent factors. 43–45
- K^{bg} Background topics. 44
- K^{fg} Foreground classes. 43, 44
- N_j Maximum number of patches/super-pixels/visual-word in image j . 45, 76, 77
- O Observed variables. 46
- T Heat map. 107, 108
- T^{bg} Index set of background topics. 44, 45
- T^{fg} Index set of foreground topics. 44, 45
- $Uniform$ Uniform distribution. 45, 46
- V_f Size of appearance vocabulary of each feature f . 44
- Φ Spatial MRF potential. 78
- Π Model parameters. 46, 79, 105

- Ψ Digamma function. 48, 80
- Θ Factorial MRF potential. 79
- β Coupling strength parameter of the MRF. 78, 105
- I Identity matrix. 76
- μ Prior mean. 105
- W_k^0 Scale matrix of Wishart distribution for topic k . 45, 46
- Λ_k^0 Prior precision of multivariate normal distribution for topic k . 45
- Λ_{kj} Precision of location distribution for topic k in image j . 45
- α Annotation/label/topic prior. 76, 82
- α_j Prior parameter in image j . 45
- α_j^{bg} Prior knowledge for background classes in image j . 45
- α_j^{fg} Prior knowledge for foreground classes in image j . 45
- μ_{kj} Mean of location distribution for topic k in image j . 45
- μ_k^0 Prior mean of multivariate normal distribution for topic k . 45
- π_{kf}^0 Appearance dirichlet distribution of topic k for feature type f . 45
- π_{kf} Appearance dirichlet distribution of topic k for feature type f . 45
- θ_j Topic distribution of image j . 45
- ϵ Small constant. 52
- A Appearance normal distribution. 76
- I Indicator function. 48, 78
- X Feature representation in a matrix form. 77
- Y Topic or factor estimation in a matrix form. 77, 78

- l_{ij} Topic location for visual word i in image j . 45
- \mathbf{x}_{jf} Feature representation of feature type f in image j . 44
- \mathcal{M} Similarity/Correlation Matrix. 52, 79
- \mathcal{NW} Normal-Wishart distribution (a conjugate prior distribution for a multivariate Gaussian). 45, 46
- \mathcal{N} Normal distribution. 45, 46, 76, 77, 105
- \mathcal{S} Student-t distribution. 48
- Bern** Bernoulli distribution. 77
- Beta** Beta distribution. 77
- Dir** Dirichlet distribution. 45, 46
- Multi** Multinomial distribution. 45, 46
- ν Variational parameter for \mathbf{y} . 80
- ρ Variable controlling the important of the factorial MRF. 79
- σ Prior variance. 76, 105
- τ Variational parameter for \mathbf{b} . 80
- θ Image prior or Topic prior. 77, 83
- φ Variational parameter for \mathbf{A} . 80
- ϑ Expected sparsity prior of annotations. 77, 105
- ξ_k^0 Degree of freedom of Wishart distribution for topic k . 45
- b Random variable of Beta distribution. 77
- f A feature type. 44, 45
- $h(\cdot)$ Histogram. 52

i Each patch/super-pixel/visual-word. [45](#), [76](#), [77](#), [83](#), [104](#), [107](#)

j Each image. [45](#), [74](#), [76](#), [78](#), [82](#), [104](#)

k Each topic/factor. [44](#), [45](#), [76](#), [83](#), [105](#), [106](#)

v Variable of vocabulary. [48](#)

y Topic or factor estimation. [77](#)

y_{ij} Topic estimation for visual word i in image j . [45](#)

Chapter 1

Introduction

Humans are extremely good at perceiving and recognising objects [7, 8] in challenging real-world scenes. A four-year-old child is a freshman of this diverse world and still has a lot to learn in his life. However, he is already an expert at one very important task: to make sense of what he sees. The child can generate the following descriptive text for the given image as Figure 1.1: *“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant.”* [1]. Over the last few decades, the computer vision community has endeavoured to imitate humans’ ability to discover and understand image content. An ideal visual system should recognise and understand the complex visual word rapidly, accurately and comprehensively. Designing an intelligent visual understanding model underpins a wide range of applications from Robotics [9], consumer photography [10] to video surveillance [11], e-commerce [12]. Figure 1.2 illustrates how an ideal intelligent vision system understands a real world image.

This thesis proposes a unified framework for automatic recognition, detection, and segmentation of objects in an image. Precisely, given an image, the goal of my research work is to answer the following questions: 1) what objects are in the image? 2) where are they? 3) what do they look like? In other words, the system aims to generate a text description containing objects, their associate attributes, together with the exact location.



Figure 1.1: An example of real world images can be easily described by a young child. [1].

1.1 Describing Images with Objects and Attributes

1.1.1 Object Recognition and Localisation

The contents of an image can be loosely categorised as belonging to *Things* or *Stuff* [13]. *Things* can be described as individual objects which have a specific size and shape (e.g., bird, car, bus, person, etc.). *Stuff* can be described as an amorphous material which defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific or distinctive spatial extent or shape (e.g., trees, water, grass, road, etc.).

Object recognition: Recognising object [14, 15] is the first step to describe the contents of an image. The image annotation task aims to identify what objects are in the image, but it does not care about the localisation of objects. When one single object dominates the whole image, there is no need to localise object anymore. The problem is then converted to image classification task. The only question people concern is what object is in the image. However, in most natural images, objects usually occupy a small portion of an image and many different objects co-exist in the same image. That

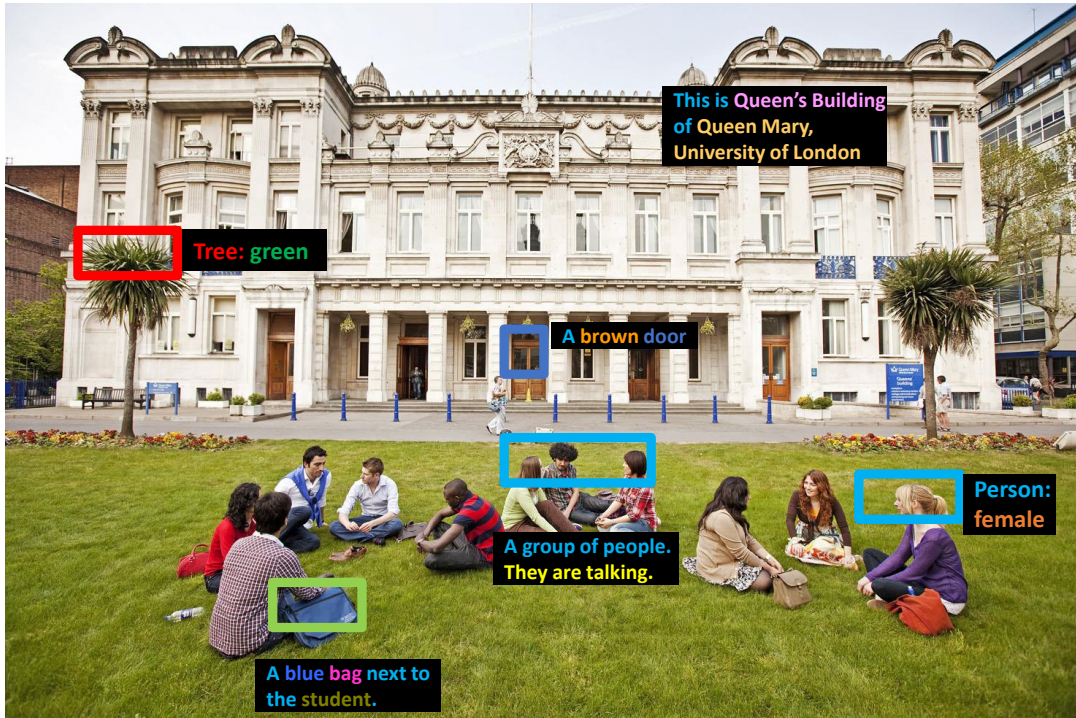


Figure 1.2: Visual understanding in real world scenes. Given a real-world image, an ideal intelligent vision system can recognise the objects within an image, such as a tree, a person. It can further describe these objects in detail by their characteristics. For example, a green tree, a group of people who are talking with each other. Moreover, it is able to identify the landmark (i.e. This is Queen's building of Queen Mary, University of London).

is why the localising object is important.

Object detection: Object detection is one of the fundamental challenges in computer vision [16]. It typically concerns the problem of detecting and localising objects from images or videos [17]. Given an image, the ideal object detection system can automatically localise objects in categories of interest. Figure 1.3 shows an example of object detection [2]. The success of early detection methods starts from localising constrained object categories, such as pedestrian or face. Recent approaches are moving the focus to the detection of varying categories with large appearance variations, such as the twenty categories of Pascal VOC [18] and one thousand categories of ImageNet [19].

Object segmentation: Object detection aims to localise and recognise every instance

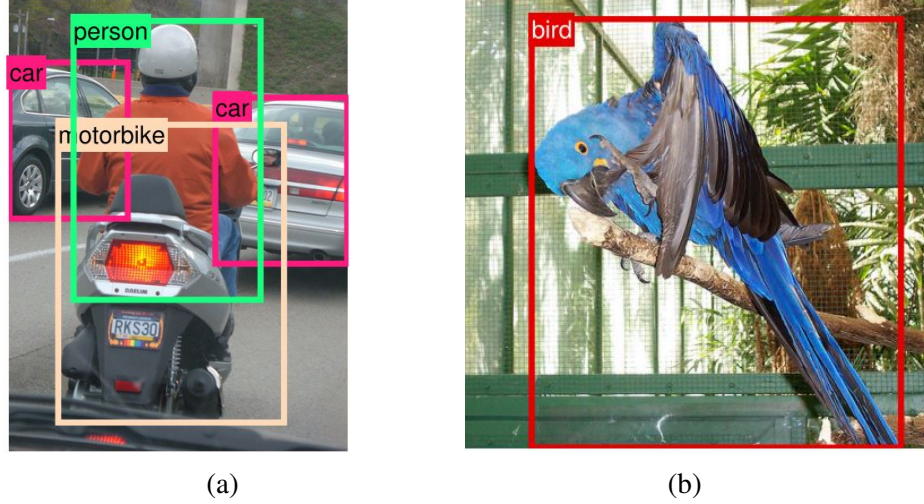


Figure 1.3: An example of object detection results from trained detector [2].

marked by a bounding box. However, bounding boxes can only provide coarse positions of detected objects. On the other hand, object segmentation [20, 21] assigns a category label to each pixel in an image, which provides more accurate locations of objects. Figure 1.4 shows an example of object segmentation [3]. Object segmentation can be considered as a pixel level classification problem. Given an image, the system aims to find the region of the object from the background. Object segmentation is a harder task. Obviously, if objects can be satisfactorily segmented, the object detection problem is thus already solved.

1.1.2 Describing Objects with Attributes

Attributes: An object also has many other qualities apart from its category. A car can be red; a shirt can be striped; a ball is round; a building is tall. These visual attributes are important for understanding object appearance and for describing objects to other people. Figure 1.5 shows some examples of attributes [4]. Most existing methods learn visual attributes from manually annotated images. Farhadi *et al.* [22] learns a broad set of complex attributes in a fully supervised manner, which assumes the bounding boxes (i.e. object localisations) are provided. Hanwell and Mirmehdi [23] also demonstrate that visual attributes can be learned in a weakly supervised setting, where training data



Figure 1.4: An example of object segmentation results from trained model [3].

are directly collected from a web image search engine such as Google Image search.

Attributes prediction: Attributes play an significant role in object representation and recognition. Thus predicting attributes accurately in given image or object is a crucial task. Attribute prediction task can be considered in two perspectives. One is learning a mapping between visual appearances and attributes. The attribute can be predicted based on the learned mapping. The other is learning the association between objects and attributes. Same objects usually contain the similar attributes. Attribute prediction is also related to object recognition and localisation. With an accurate recognition and localisation of an object, the model can predict attributes better by observing a clean appearance. At the same time, the better attribute prediction can also help the system to recognise and localise object. These two tasks are closely connected and help each other.

Attribute as representation for other vision tasks: Attribute based approach allows many related tasks which cannot be performed before. To effectively recognise object categories is not the only benefit. The system can also describe unknown object categories, report atypical attributes of known classes, and even learn models of new object categories from a purely textual description. The attribute can further extend object recognition into fine-grained or instance level. Different objects from the same category may have a subtle difference. Describing objects by attributes can distinguish



Figure 1.5: Examples of different kinds of attributes [4].

them more easily.

Describing images with objects and attributes: Vision research is moving beyond simple classification, annotation and detection to encompass a more complete image understanding task, such as generating more structured and semantic descriptions of images. When humans describe images, using visual attributes is a semantically rich way to distinguish objects in the world. More specifically, a sentence to describe images usually uses combinations of objects and their associated attributes. For example, an image can be described as containing “a person in red clothes and a shiny car”. In order to imitate this ability, a computer vision system needs to learn models of objects, attributes, and their associations.

1.1.3 Applications

Describing an object with attributes facilitate a variety of vision tasks including recognition, classification, image description and retrieval. Addressing these tasks will underpin a wide range of applications from robotics [9], pilotless automobile [24], consumer photography [10] to video surveillance [11], gaming [25], e-commerce [12]. The following gives a brief introduction of some applications that related to the proposed approach. These applications are changing the way people live.

Image caption generator Visual understanding is the ability that humans use to discover and analyse the surroundings. Humans usually can point out and describe the



Figure 1.6: An example of generating detailed description of images [5].

details of each part of an image with only a quick glance. Generating a detailed description [5, 26] for given images is a highly desired property of computer. This will facilitate to efficient organising and indexing multimedia data, especially in the era of data explosion recently. Figure 1.6 shows an example of generating detailed description of image [5].

Image search: Humans often have very specific visual content in mind about what kind of objects they are searching for. The most natural way for people to communicate their target object is to describe it in terms of its attributes [27]. For example, given the query “young Asian woman wearing sunglasses”, the designed system aim to infer that relevant images are likely to have all these attributes. Figure 1.7 shows an example of image search given a multi-attribute query [6]. The ability of current search engines to find images still heavily rely on text annotations attached to images. The interested images are searched based on matching text descriptions. In fact, the annotations of images are often inaccurate or confused as they may refer to other related content.

Video surveillance: Surveillance is generally used to monitor the behaviour, activities, or other changing information. It is typically achieved by the observation from a distance using electronic equipment (such as CCTV cameras). The main target of cameras is people, as well as the related behaviour including influencing, managing, directing,

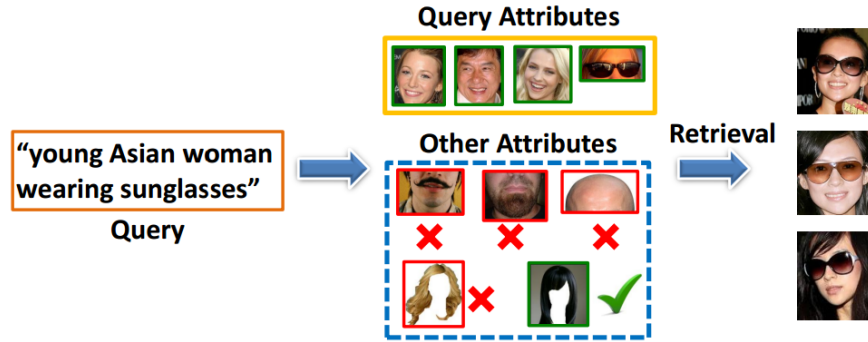


Figure 1.7: An example of image search given a multi-attribute query [6].

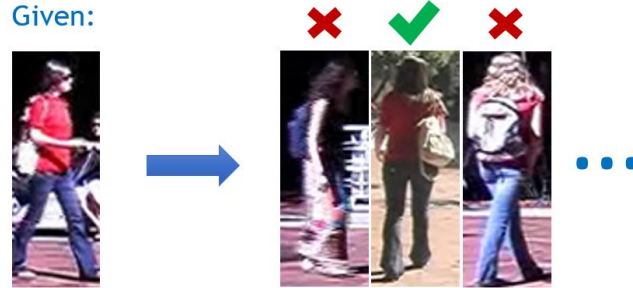


Figure 1.8: An illustration of person re-identification

or protecting them. Person Re-identification (Re-ID) and description-based search are crucial tasks in visual surveillance. They underpin many fundamental applications including multi-camera tracking, crowd analysis and forensic search. Both tasks aim to retrieve images of a specific person but differ in the query used. **Person Re-ID** (see Figure 1.8) queries using an image from a different view (e.g., in multi-camera tracking), while **person search** (see Figure 1.9) uses a textual person description (e.g., eyewitness description).



Figure 1.9: An illustration of person search

1.2 Challenges and Motivations

1.2.1 General Challenges for Object and Attribute Recognition

Object recognition is a challenging problem especially on a large scale. These challenges broadly fall into three categories: 1) Objects are easily occluded or confused by background clutter. It is very common that backgrounds may share the same colour or shape with the object of interest. It is difficult to separate foreground and background especially when objects exist in a noisy background. 2) The object itself can appear with the great variabilities in object appearance, viewpoint, illumination and pose [28, 29, 30]. The appearance of the same type of object or even exactly the same object can easily change when the environmental condition varies. For example, “blue” can be recognised to “purple” when the lighting condition change. 3) Objects in different domains also cause an appearance variation. For example, typical object detectors trained on images perform poorly on video, as there is a clear distinction in the domain between the two types of data. This phenomenon is known as “domain shift”. Specifically, simply applying the classifier learned in the source domain may hurt the performance in the target domain. It is even harder to annotate data when the surveillance scenario is considered. The data captured by low-resolution camera suffer various poor condition including the variability of viewpoints, illumination, pose, partial occlusion and motion-blur [31].

1.2.2 Challenges Faced by Fully Supervised Learning based Approaches

In order to separate foreground object from background clutter, fully/strongly annotated data are thus typically required to learn a generalisable model for tasks such as object classification [32], detection [16, 33], and segmentation [30, 34, 35]. In fully annotated images, such as those in the PASCAL VOC object classification or detection challenges [18], not only the presence of objects, but also their locations are labelled, typically in the form of bounding boxes. Most conventional methods for Weakly Supervised Object Localisation (WSOL) learn a discriminative based model on these fully labelled data. They typically treat different label/object category independently during the learning process. A classic method is to learn a binary classifier to distinguish between the object of interest and all other object.

When moving from object to object with attributes, the annotation needs to be more detailed. Specifically, in conventional pipeline images are strongly labelled with object bounding boxes or segmentation masks, and associated attributes, from which object detectors and attribute classifiers are trained. Given a new image, the learned object detectors are first applied to find object locations, where the attribute classifiers are then applied to produce the object descriptions.

However, the scalability issue prevents it being applied to large scale data. Considering there are over 30,000 object classes distinguishable to humans [36], the equally large number of attributes to describe them [37], and much larger number of combinations of objects and attributes, Fully Supervised Learning (FSL) is not scalable due to the lack of fully labelled training data. Thus, it is clear to see that a larger difficulty of obtaining sufficient and sufficiently detailed annotations to learn robust and accurate object or attribute detectors.

1.2.3 Challenges Faced by Weakly Supervised Learning based Approaches

Manual annotation of hundreds of object categories is time-consuming, laborious, and subjective to human bias. Media data are increasingly available with the prevalence of sharing websites such as Flickr, however, the lack of annotated images, particularly

strongly annotated ones, becomes the new barrier that prevents tasks such as object detection from scaling to thousands of classes [38].

Weakly Supervised Learning (WSL) aim to model complex visual scenes from weakly labelled images abundant on media sharing sites. These methods aim to learn the appearance of objects and attribute classes as well as their associations. WSL methods largely reduced extensive manual annotation compare with conventional FSL methods on computer vision tasks. Although WSL solved the scalability issue, however, without accurate annotation, a new challenge has been introduced: ambiguity.

WSL method are a desired way to reduce the amount of manual annotation. One natural image is typically attached with several tags, which are image-level labels (either object or attributes) without their locations and associations. However, learning strong semantics, i.e. explicit object-attribute association for each object instance from weakly labelled images, is extremely challenging due to the label ambiguity: a real-world image with the tags “dog, white, coat, furry” could contain a furry dog and a white coat or a furry coat and a white dog. Furthermore, the tags/labels typically only describe the foreground/objects. There could be a white building in the background which is ignored by the annotator, and a computer vision model must infer that this is not what the tag ‘white’ refers to. A desirable model thus needs to jointly learn multiple objects, attributes and background clutter in a single framework in order to explain away ambiguities in each by knowledge of the other. Moreover, there are a potentially unlimited number of attributes co-existing on a single object (e.g. there are hundreds of different ways to describe the appearance of a person) which are almost certainly not labelled exhaustively in each training image. They also need to be modelled so that they do not act as distractors that have a detrimental effect on the understanding of objects and attributes of interest. For instance, if annotators only labelled a banana in a training image but did not bother to label it as yellow. Even if yellow has never been used as an attribute label for any object in the training set, the model should be able to infer yellow as a latent attribute and associate it with the banana, so that other colours would not be assigned wrongly.

1.2.4 Motivations

Based on above discussion, the work is motivated by three factors:

1) Existing methods for recognition tasks excel by their capacity to take advantage of massive amounts of fully supervised training data. This reliance on full supervision is a major limitation on scalability with respect to the number of classes or tasks. In order to solve the scalability issue, WSL is preferred. As motivated by other WS methods, the proposed approach aims to exploit limited weak label information. Weakly labelled data are widely available in the era of data explosions compare to the manually labelled strong annotation. Therefore, designing a framework to directly utilise image level weak tags is the main motivation of this thesis.

2) WSL has to face the additional challenge of label ambiguity. This is difficult to learn a clean model without annotation information of correspondence among object labels, attributes and pixels, involving the relation such as which category label correspond to which pixel, which attribute label correspond to which category label, etc. Most existing WS approaches are based on Independent Learning (IL) manner. They ignore the fact that many object categories share some common property. This thesis argues that IL loses the information shared among classes. WSL should fully use all the cues that available, because the information is already very limited. This thesis aims to design a joint learning framework to model all classes together, so as to explore the knowledge of each class and the shared properties among them. Beyond the relation between each class, joint learning based approach can also model attributes and background together with objects.

3) Most existing approaches for WSOL are based on a discriminative model. However, the generative based approach enables us to exploit a number of additional knowledge including prior information and cues transferred from other auxiliary data. This is important to help reducing the ambiguity of learning object appearance especially in this WS scenario. The capacity of transfer learning is another advantage of generative based approach. This allows us to learn reliable knowledge from existing auxiliary dataset (sometimes even fully labelled data) and adapt it to help learning target weakly labelled data.

1.3 The Proposed Approach

1.3.1 A Joint Bayesian Approach for Weakly Supervised Learning

In this thesis, a generative approach is proposed to address the challenging of WSL. There are mainly three general advantages of generative based approach to deal with the ambiguity:

Joint vs. independent modelling By jointly modelling different classes of objects and background, the proposed approach is able to exploit multiple object co-occurrence, so each object known to appear in an image can help disambiguate the location of the others by accounting for some of the pixels. Meanwhile, a single set of shared background topics is learned once for all object classes. This is due to the nature of a generative model – every pixel in the image must be accounted for. Even though learning background appearance can further disambiguate the location of objects, this appears to be an extremely hard task given that no labels are provided regarding background (people tend to focus on the foreground when annotating an image). However, by learning them jointly with the foreground objects and using all training images available, this task can be fulfilled effectively by the proposed models.

Latent graphical model The proposed approach is based on a latent graphical model. The key advantage of addressing WSL by a latent graphical model is that it can naturally model the distribution of all objects and backgrounds together. It is unlike discriminative based method which aims to learn the hard or soft boundary between objects as well as between the object and background individually. A latent graphical based model can easily add the observed variable to learn the appearance of the object of interest. Latent variable can also be introduced to discover the hidden information or new things. This intuitive explanation ability is irreplaceable. Another motivation for using latent graphical model is to utilise related prior knowledge flexibly. For example, both external human or internal data-driven prior about typical object size, location and appearance could be considered as a Bayesian prior. This is particularly important in WS setting that only limited information available.

Bayesian domain adaptation Elaborate labelling of object and attributes for every individual domain/dataset is time-consuming and expensive. In fact, some objects and attributes are shared or related to a group of domains. It performs poorly if the

learned model is directly applied to a different domain [39]. Thus, domain adaptation is usually required to bridge the gap. A central challenge for building generally useful recognition models is providing the capability to adapt models trained on one domain or dataset to new domains or datasets [40]. This is important because any given domain or dataset is intentionally or unintentionally biased [41], so transferring models directly across domains generally performs poorly [41]. However, with appropriate adaptation, source and target domain data can be combined to out-perform target domain data alone [40]. The designed model's Bayesian formulation enable it to provide domain adaptation in computer vision tasks.

The proposed approach has been taken to develop the following three models:

1.3.2 Weakly Supervised Localisation of Objects

First, this thesis addresses the problem of localisation of objects as bounding boxes in images and videos with weak labels (see Figure 1.10). This WSOL problem has been tackled in the past using discriminative models where each object class is localised independently from other classes. In this thesis, a novel framework based on Bayesian joint topic modelling is proposed, which differs significantly from the existing ones in that: (1) All foreground object classes are modelled jointly in a single generative model that encodes multiple object co-existence so that “explaining away” inference can resolve ambiguity and lead to better learning and localisation. (2) Image backgrounds are shared across classes to better learn varying surroundings and “push out” objects of interest. (3) The proposed model can be learned with a mixture of weakly labelled and unlabelled data, allowing the large volume of unlabelled images on the Internet to be exploited for learning. Moreover, the Bayesian formulation enables the exploitation of various types of prior knowledge to compensate for the limited supervision offered by weakly labelled data, as well as Bayesian domain adaptation for transfer learning. Extensive experiments on the PASCAL VOC, ImageNet and YouTube-Object (YTO) videos datasets demonstrate the effectiveness of the proposed Bayesian joint model for WSOL.

One approach to this challenge is WSOL: simultaneously locating objects in images and learning their appearance using only weak labels indicating presence/absence of the objects of interest. The WSOL problem has been tackled using various ap-

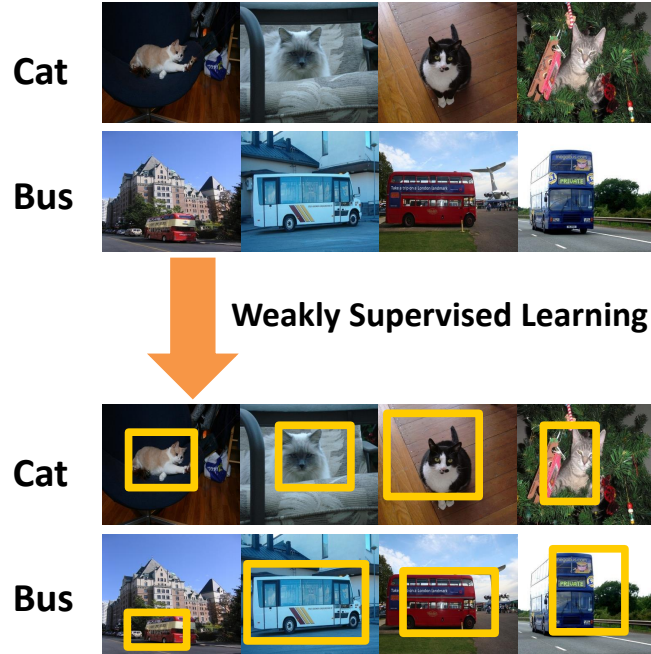


Figure 1.10: Weakly supervised object localisation

proaches [32, 38, 42, 43, 44, 45, 46]. Most of them address the task as a WSL problem, particularly as a Multi-instance Learning (MIL) problem, where images are bags, and potential object locations are instances. These methods are typically discriminative in nature and attempt to localise each class of objects independently from the other classes. However, localising objects of different classes independently has a number of limitations: (1) It fails to exploit the knowledge that different objects often co-exist within an image. For instance, knowing that some images have both a horse and a person, in conjunction with a joint model for both classes – the person can be “explained away” to reduce ambiguity about the horse’s appearance, and vice versa. Ignoring this increases ambiguity for each class. (2) Although object classes vary in appearance, the background appearance is relevant to them all (e.g. sky, tree, and grass are constant features of an image regardless of the foreground object classes). When different classes are modelled independently, the background must be re-learned repeatedly for each class, when it would be more statistically robust [47] to share this common knowledge.

In this thesis, a novel framework based on Bayesian Latent Topic Model (LTM) is proposed to overcome the mentioned limitations. In the presented framework, both multiple object classes and background types are modelled jointly in a single generative model as latent topics, in order to explicitly exploit their co-existence relationship. As Bag-of-Words (BoW) models, conventional LTMs have no notion of localisation. The proposed model overcomes this problem by incorporating an explicit notion of object location.

Apart from joint learning and domain transfer, the designed generative model based framework has the following advantages over previous discriminative approaches:

Integration of prior knowledge Exploiting prior knowledge or top-down cues about appearance or geometry (e.g., position, size, aspect ratio) should be supported if available to offset the weak labels. The proposed framework is able to incorporate, when available, prior knowledge about appearances of objects in a more systematic way as a Bayesian prior. Going beyond within-class priors, the designed method also shows that cross-class appearance similarity can be exploited. For instance, the model can exploit the fact that “bike” is more similar to “motorbike” than “aeroplane”.

Semi-supervised Learning (SSL) Since there are effectively unlimited quantity of unlabelled data available on the Internet (compared to a limited quantity of manually annotated data), a valuable capability is to exploit this existing unlabelled data in conjunction with limited weakly labelled data to improve learning. As a generative model, the proposed framework is naturally suited for SSL. Unlabelled data are included and the label variables for these instances left unclamped (i.e. no supervision is enforced). Importantly, unlike conventional SSL approaches [48], the presented model does not require that all the unlabelled data are instances of known classes, making it more applicable to realistic SSL applications.

1.3.3 Weakly Supervised Learning of Objects and Attributes

Second, this thesis proposes to model complex visual scenes using a non-parametric Bayesian model learned from weakly labelled images abundant on media sharing sites such as Flickr. Given weak image-level annotations of objects and attributes without their locations and associations between them, the designed model aims to learn the appearance of objects and attribute classes as well as their associations on each object

instance. Once learned, given an image, the proposed model can be deployed to tackle a number of vision problems in a joint and coherent manner, including recognising various objects in the scene (object annotation), describing the objects using their attributes (attribute prediction and association), and localising and delineating the objects (object detection and semantic segmentation). This is achieved by developing a novel Weakly Supervised Markov Random Field Stacked Indian Buffet Process (WS-MRF-SIBP) that models object and attributes as latent factors and explicitly captures their correlations at both image and superpixel levels. Extensive experiments on benchmark datasets demonstrate that the proposed WS model significantly outperforms alternative WS alternatives and is often comparably with existing strongly supervised models on a variety of tasks including semantic segmentation, image annotation and retrieval based on object-attribute associations.

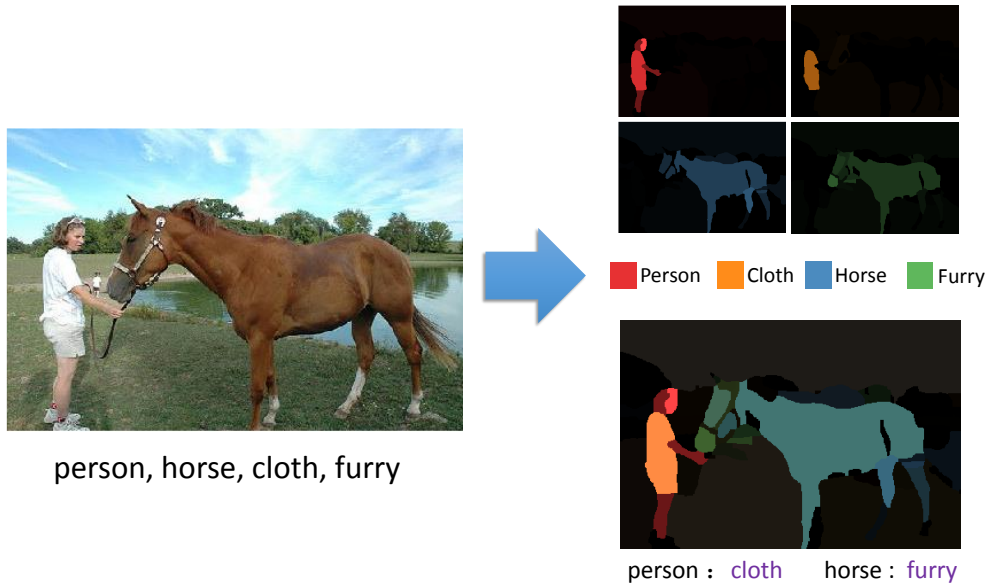


Figure 1.11: Weakly supervised learning of objects and attributes

A novel unified framework is developed which capable of jointly learning objects, attributes and their associations. Figure 1.11 shows an example of the input and output of our system, where weak annotation in the form of a mixture of objects and attributes is transformed into objects and attributes associations with object segmentation. Under the framework, given a training image with image level labels of objects and attributes,

the image is first over-segmented into superpixels; the joint object and attribute annotation and segmentation problem thus becomes a superpixel multi-label classification problem whereby each superpixel can only have one object label but an arbitrary number of attribute labels. Treating each label as a factor, this work develops a novel factor analysis model by generalising the non-parametric Indian Buffet Process (IBP) [49]. The IBP is chosen because it is designed for explaining multiple factors that simultaneously co-exist to account for the appearance of a particular image or superpixel, e.g., such factors can be an object and its particular texture and colour attributes. Importantly, as an infinite factor model, it can automatically discover and model latent factors not defined by the provided training data labels, corresponding to latent object/attributes as well as structured background clusters (e.g. sky, road). However, the conventional IBP is limited in that it is unsupervised and, as a flat model, applies to either superpixels or images, not both; it thus cannot be directly applied to the current problem. Furthermore, the standard IBP is unable to exploit cues critical for segmentation and object-attribute association by modelling the correlation of factors within and across superpixels in each image. Specifically, the within-superpixel correlation captures the co-occurrence relations such as cars are typically metal and bananas are typically yellow, whilst the across-superpixel correlation dictates that neighbouring superpixels are likely to have similar labels. To overcome these limitations, a novel variant of IBP, termed WS-MRF-SIBP is formulated in this thesis. It differs from the conventional IBP in the following aspects: (1) By introducing hierarchy into IBP, WS-MRF-SIBP is able to group data, thus allowing it to explain images as groups of superpixels, each of which has an inferred multi-label description vector corresponding to an object and its associated attributes. (2) It learns from weak image-level supervision, which is disambiguated into multi-label superpixel explanations. (3) Two types of Markov Random Field (MRF) over the hidden factors of an image are introduced to the model hierarchy: across-superpixel MRF to exploit spatial coherence among neighbouring superpixels and within-superpixel MRF to exploit co-occurrence statistics of different factors within a given superpixel.

1.3.4 Transferring Learning of Objects and Attributes

Finally, this thesis tackles the WSL problem from a transfer learning perspective. Learning semantic attributes for describing person search has gained increasing interest due to attributes' great potential as a pose and view-invariant representation. However, existing attribute-centric approaches have thus far underperformed state-of-the-art conventional low-level features based approaches. This is due to their non-scalable need for an extensive domain (camera) specific annotation. This thesis presents a new semantic attribute learning approach for describing a person. The proposed model is trained on existing fashion photography datasets – either weakly or strongly labelled (see Figure 1.12). It can then be transferred and adapted to provide a powerful semantic description for surveillance person detections, without requiring any surveillance domain supervision. The resulting representation is useful for both unsupervised and supervised person Re-ID applications, achieving state-of-the-art and near state-of-the-art performance respectively. Furthermore, as a semantic representation it allows description-based person search to be integrated into the same framework.

Person Re-ID and description-based search are crucial tasks in visual surveillance. Although extensive research [50, 51] have been conducted for these tasks, it still remains unsolved due to various challenges including the variability of viewpoints, illumination, pose, partial occlusion, low-resolution and motion-blur [31].

Despite their hoped-for potentials, attribute-centric approaches to person Re-ID have until now not achieved state-of-the-art performance compared to conventional alternatives focused on learning effective low-level features and matching models [51]. This thesis argues that this is largely due to the difficulty of obtaining sufficient and sufficiently detailed annotations to learn robust and accurate attribute detectors. In particular, to achieve good attribute detection, per camera/dataset annotations need to be obtained, since attribute models typically do not generalise well across cameras (e.g. a blue shirt may look purple in a different camera view). This is exacerbated, because unlike annotation of person identity used in learning a Re-ID matching model, attributes require multiple labels per image. Moreover, since most human attributes are associated with a specific body part, to learn accurate attribute detectors, ideally annotation needs to be done at the patch-level rather than the image-level. In short, an attribute-based approach is limited by the scale of its annotation requirements.

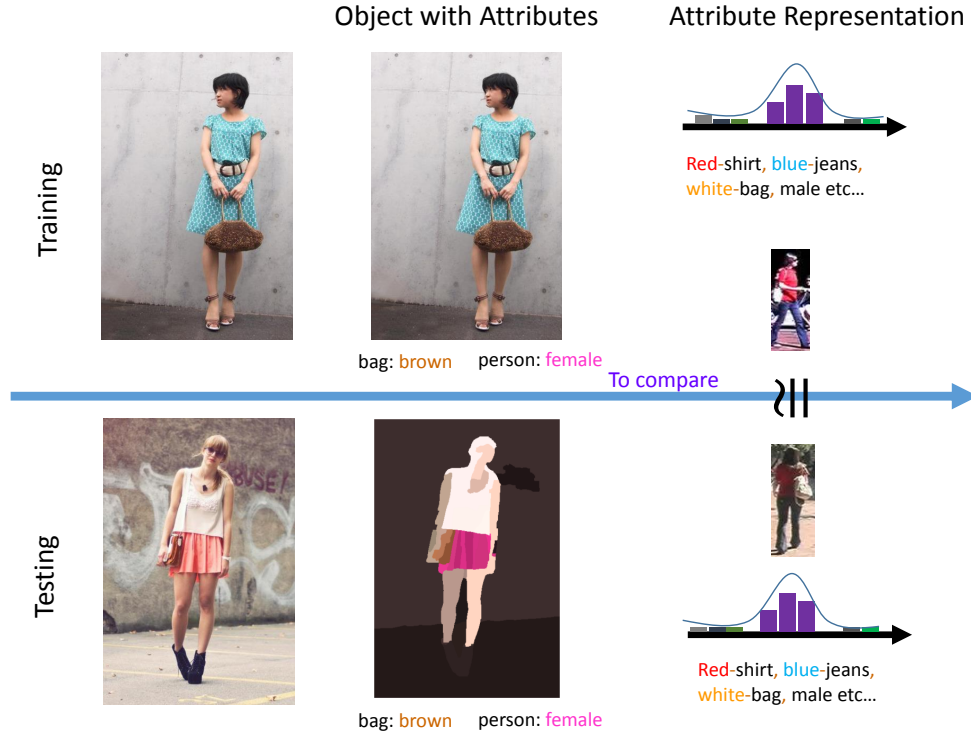


Figure 1.12: Transferring a Semantic Representation

Although attribute-centric approaches to surveillance are limited by the lack of annotation, extensive person attribute annotations already exist in other domains, notably fashion image analysis [52, 53, 54, 55, 56]. In this line of work, many high resolution images have been annotated with clothing properties – sometimes strongly (per-pixel). However, learning attribute detectors from the fashion domain, and applying them directly to person Re-ID and search will fail, due to the severe domain shift problem – compared with surveillance data, the image characteristics are completely different (see Figure 5.4). In particular, many of the challenges in surveillance are absent (e.g. illumination variability, occlusion, low-resolution, motion blur). These large and perfectly annotated person attribute datasets are thus useless, unless an attribute model learned from them can be successfully adapted and transferred to the surveillance domain.

In this thesis, a new framework is contributed that is capable of learning a semantic attribute model from existing fashion datasets, and adapting the resulting model

to facilitate person Re-ID and search in the surveillance domain. In contrast to most existing approaches to attribute detection [57, 58] which are based on discriminative modelling, this thesis takes a generative modelling approach based on the IBP [49]. The generative formulation provides key advantages including: joint learning of all attributes; ability to naturally exploit weakly-annotated (image-level) training data; as well as unsupervised domain adaptation through Bayesian priors. Importantly an IBP-based model [59, 60, 61] provides the favourable property of combining attributes factorially in each local patch. This means that the presented model can differentiate potentially ambiguous situations such as Red-Shirt+Blue-Jeans versus Red-Jeans+Blue-Shirt. Moreover, with this representation, attribute combinations that were rare or unseen at training time can be recognised at test time so long as they are individually known (e.g. Shiny-Yellow-Jeans).

The proposed framework overcomes the significant problem of domain shift between fashion and surveillance data in an unsupervised way by Bayesian adaptation. It can exploit both strongly and weakly annotated source data during training, but is always able to produce a strong (patch-level) attribute prediction during testing. The resulting representation is highly person variant while being view-invariant, making it ideal for person Re-ID, where the proposed model obtains state-of-the-art results. Moreover, as the representation is semantic (nameable or describable by a human), the presented method is able to unify description based person search within the same framework, where it also achieve state-of-the-art results.

1.4 Contributions

This thesis makes the following key contributions:

1. This thesis first formulates a novel Bayesian topic model suitable for WSOL, which can use various types of prior knowledge including an inter-category appearance similarity prior. The Bayesian prior also enables the model to easily borrow available domain knowledge from existing auxiliary datasets and adapt it to a target domain.
2. This thesis further jointly learns all object, attribute and background appearances, object-attribute association, and their locations from realistic weakly la-

belled images including multiple objects with a cluttered background. To this end, a novel WS non-parametric Bayesian model is formulated by generalising the IBP to make it WS, hierarchical, and integrated with two types of MRFs over hidden factors to both learn and exploit spatial coherence and factor co-occurrence.

3. A generative framework is introduced for person attribute learning that can be learned from strongly or weakly annotated data or a mix. The designed approach shows how to perform domain adaptation from fashion to surveillance domain.
4. Once learned from weakly labelled data, the proposed models can be deployed for various tasks including object localisation, semantic segmentation, image description and image query, many of which rely on predicting strong object-attribute association. Extensive experiments on benchmark datasets demonstrate that on all tasks the presented model significantly outperforms a number of WS baselines and in many cases is comparable to the strongly supervised alternatives.

1.5 Outline

This thesis is organised into seven chapters:

Chapter 2 presents a review of various existing strategies and approaches for object recognition and complex visual understanding, while providing further motivations for the proposed approaches of this thesis.

Chapter 3 provides detailed explanations on the the proposed Bayesian LTM for WSOL.

Chapter 4 presents the Weakly-supervised Stacked Indian Buffet Process (WS-SIBP) for modelling objects and attributes to discover the associations between them.

Chapter 5 explains how the proposed model transfer semantic attributes to provide a powerful representation for describing objects.

Chapter 6 provides conclusions and suggests a number of areas to be pursued as further work.

Chapter 2

Literature Review

The work presented in this thesis is related to a wide range of computer vision problems including image classification, object recognition, attribute learning, and semantic segmentation. In this chapter, some of most closely related topics are reviewed. First, this review briefly looks at object detection and segmentation in Section 2.1 and Section 2.2 respectively. Both of them are discussed from Fully Supervised (FS) methods to WS ones. Then, Section 2.3 focus on existing approaches to attribute learning, involving attribute prediction and attribute representation for other vision tasks. Following that, a summarisation of state-of-the-art methods on joint modelling objects and attributes is provided in Section 2.4. Existing and related WSL techniques are described in Section 2.5. Finally, the review is summarised in Section 2.6.

2.1 Object Detection

2.1.1 Fully Supervised Object Detection

Over the past few years, researchers have made significant progress in object detection. Most of these efforts have generally followed two directions. One line of work has focused on proposal generation techniques with the goal of efficient and accurate localising candidate objects. These methods aim to achieve high detection recall with as few proposals as possible. A second line of work has explored the use of learning techniques for classifying proposals. The task is to assign a label of foreground class or background to each possible proposals.

Objects can appear at a variety of locations and sizes within an image. The dominant approach to object detection tasks has been the sliding windows fashion. It is an exhaustive search procedure for all different sizes placed at all locations. Sliding window paradigm has been widely adopted by pioneer works in face detection [62] and pedestrian detection [63]. To date, it is still a common strategy for recent state-of-the-art pedestrian detection [64] and face detection [65]. The Deformable Part Model (DPM) is a well-known champion systems of object detection in Pascal VOC 2007-2011 challenge, which is also based on sliding window. The main drawback of the sliding window is that the number of candidate proposals can be about $O(10^6)$ for a 640×480 image. In the work of Yan *et al.* [66], a large number of windows are required to be scored in a test image using a classifier. Heavy computational load severely limits the applicability for real time detection due to the evaluation inefficiency.

To keep the computational cost manageable, object proposal gains attentions to reduce the number of evaluated windows. Previous works demonstrate that the proposal algorithm can indeed significantly speed up the object detection task and improves its performance. Object proposal has been proven to be very useful for object detection [67]. It has been popularised recently [68, 69] as a powerful pre-processing approach for computational efficiency and accuracy. Uijlings *et al.* [68] propose object proposals by grouping low-level superpixels from [70] hierarchically. The number of proposals can be about 2, 000 with a recall rate of 98% on PascalVOC and 92% on ImageNet. This greatly reduce the number of evaluated windows with only minor loss in detection performance. It gives object detection algorithm more flexibility to localise things of interest by providing a smaller number of arbitrary scale and aspect ratio proposals. These kinds of methods was subsequently adopted to achieve impressive object detection performance on Pascal VOC [71] and ImageNet [72]. Object proposal has been continuously improving in recent methods by generating more compact and efficient proposals. Most methods are trained by means of supervised learning [69, 73, 74, 75], while some others are based on unsupervised approach [67, 68, 76]. Hosang *et al.* [77] provide an in depth analysis of ten recent object proposal methods and discuss their common weakness as well as the insights to choose the most adequate method for different settings.

Actually, object detection is essentially a classification problem for each proposal, when the proposals are produced on testing images. The proposals are typically repre-

sented as a feature vector and then feed them into a designed classifier to distinguish the foreground object and background clutter with representation. The impressive performance has been achieved thanks to discriminative learning and carefully crafted features. The feature representation becomes more sophisticated over time with the enormous amount of efforts on analysing of color, shape, texture and motion. Recent representative methods include, but are not limited to, hand-crafted Haar [62] and HOG [78] to learning based CNN [71]. A carefully designed discriminative models is typically trained on top of these feature representations. One of the most heavily studied paradigms for object detection is the deformable part-based model [16]. Beyond low-level features, there are some works on encoding scheme, such as the BoW [79] and high order feature vectors [80, 81]. A broad number of variants of classical discriminative classifier (including Boosting [82], SVM [83]) have been proposed to improve object recognition performance. Structural SVM [84, 85] is one of the most noticeable generalisation to address the complementary issue of problems involving complex outputs such as multiple dependent output variables and structured output spaces. It has been further refined to a latent version [86] to discover the dependencies on a set of unobserved latent variables. More recently, the Convolutional Neural Network (CNN) is being widely used in object detection systems since it enjoyed a great success in large-scale visual recognition challenge [87]. One key advantage of using CNN is that it can learn features directly from the images. The automatically extracted CNN features turn out to be highly versatile and can be more effective than traditional hand-crafted features on visual computing tasks. A recent work [88] also shows that the DPM can be interpreted as a CNN. The CNN based representation has shown great potentials and has been adopted by most of the leading methods in ImageNet challenge [89].

All existing object detectors take a FSL approach, in which all the training images are manually annotated with the object location. However, manual annotation of hundreds of object categories is time-consuming, laborious, and subjective to human bias. To reduce the amount of manual annotation, WSL approach is desired.

2.1.2 Weakly Supervised Object Detection

In WSL, the training set is only annotated with a binary label indicating the presence or absence of the object of interest, not the location or extent of the object. WSL approaches first locate the object of interest in the training images and then the location information is used to train a detector in a FS fashion.

WSL has attracted increasing attention as the volume of data which interested in learning from grows much faster than available annotations. WSOL is of particular interest [32, 42, 43, 44, 46, 90, 91, 92, 93], due to the onerous demands of annotating object location information. Many studies [32, 42] have approached this task as a MIL [94, 95] problem. However, only relatively recently have localisation models capable of learning from challenging data such as the PASCAL VOC 2007 dataset been proposed [42, 43, 44, 90, 91]. Such data are especially challenging because objects may occupy only a small proportion of an image, and multiple objects may occur in each image: corresponding to a Multi-instance Multi-label (MIML) problem [96]. Three types of cues are exploited in existing WSOL approaches: (1) *saliency* – a region containing an object should look different from the majority of (background) regions. The object saliency model in [69] is widely used in most recent work [38, 40, 42, 43, 90] as a preprocessing step to propose a set of candidate object locations so that the subsequent computation is reduced to a tractable level, (2) *intra-class* – a region containing an object should look similar to the regions containing the same class of objects in other images [90], and (3) *inter-class* – the region should look dissimilar to any regions that are known to not contain the object of interest [42, 43, 44]. One of the first studies to combine the three cues for WSOL was [42] which employed a Conditional Random Field (CRF) and generic prior object knowledge learned from a fully annotated dataset. Later, Pandey and Lazebnik [44] presented a solution exploiting latent SVMs. Recent studies have explicitly examined the role of intra- and inter-class cues [43, 90], as well as transfer learning [38, 91], for this task. Similar to the above approaches for weakly labelled images, [40, 97] proposed video based frameworks to deal with motion segmented tubes instead of bounding-boxes. In contrast to these studies, which are all based on discriminative models, this thesis introduces a generative topic model based approach that exploits all three cues, as well as joint multi-label, semi-supervised and cross-domain adaptive learning.

From another perspective, early works on WS annotation mainly focused on saliency based approaches [98, 99, 100]. While these methods provided a set of potential salient object location, they were shown to perform poorly for automatic annotation of objects in challenging cluttered images [101]. Recently many methods [32, 90, 101] re-cast the WS problem as a MIL problem. In a MIL formulation, each image with the object of interest is treated as a positive bag with many instances (potential object locations) of which at least one is positive and the images without the object of interest are treated as negative bags with only negative instances. The MIL based algorithms iteratively select the positive instances in each positive bag using inter-class and/or intra-class information. The approach by Nguyen *et al.* [32] is an inter-class method that defines the entire positive image as the initial positive instance and then trains a Support Vector Machine (SVM) to separate these initial positive instances and the negative images. The trained SVM is then used as a detector on the positive training images to refine the object location. However, the initial assumption is that the entire image is a good representation of the object, which is not always true. Pandey and Lazenby [44] relaxed this assumption by using a latent SVM that treats the actual location of the object as a latent variable which is a constraint to be at least 40% overlapped with the entire image. Unlike the inter-class information based methods [32, 44], Deselaers *et al.* [101] and Siva and Xiang [90] use saliency [98] to define the initial instances in each image. Then the positive instances are iteratively selected by optimising a cost function based on both inter-class and intra-class information.

2.2 Object Segmentation

2.2.1 Fully Supervised Segmentation

One of the earliest image segmentation approaches, published more than 40 years ago by Muerle and Allen [102], aimed to compute object regions by iteratively merging similar small patches, where color and texture [103] provide important information cues. Object segmentation [21, 104, 105, 106] have witnessed tremendous progress over the last decades. Arbelaez *et al.* [105] convert the problem of image segmentation to contour detection. The output of any detected contour map can be transformed to a hierarchical region tree by its segmentation algorithm using generic machinery.

Carreira and Sminchisescu [21] present a novel framework to generate and rank plausible hypotheses for the spatial extent of objects in images using bottom-up computational processes and mid-level selection cues. The generated segmentation hypotheses can be successfully used in a segmentation-based visual object category recognition pipeline [107]. Kim and Grauman [104] introduce a category-independent shape prior for object segmentation, where shared shape knowledge is discovered between objects of different categories. Weiss and Taskar [106] propose a flexible method for object segmentation that integrates rich region-merging cues with mid- and high-level information about the object layout, class, and scale into the segmentation process.

All the above methods focus on segmenting foreground objects from the background. They are not interested in the different types of backgrounds. Semantic segmentation [108] aims to assign a category label to each pixel of images, where background is also associated with various “stuff” such as sky, road, trees, etc. The prior approaches on semantic segmentation can be broadly classified into two categories: learning-based and non-parametric models.

The conventional learning-based models observe the appearance of semantic classes under various transformations. A joint parametric model is usually adopted to relate different types of information cues. Semantic segmentation using CRF based models [109, 110, 111, 112] consistently gain improvement recently. One major advantage of CRF is that it can jointly model the structure and appearance of an image. More specifically, CRF based formulation provides a natural mechanism to combine unary potentials obtained from the visual features of super-pixels with the neighborhood constraints. Unlike these flat CRF based approaches, Lempitsky *et al.* [113] represent an image by a hierarchical segmentation tree. A joint CRF is optimised using graph cut with all the benefits of global optimality and efficiency, where the main advantage is come from the utilisation of higher order contextual information. More recently, Farabet *et al.* [114] present a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centred on each pixel. The method alleviates the need for engineered features, and produces a powerful representation that captures the texture, shape, and contextual information. Pinheiro and Collobert [115] has further extended it by considering a recurrent convolutional neural network which can take into account long range label dependencies in the scenes while controlling the capacity of the network.

On the other hand, non-parametric approaches have received increasing attention [116, 117, 118, 119, 120, 121] for semantic segmentation (also known as natural scene parsing). They typically guide the segmentation through a sophisticated template matching method to retrieve images with similar visual appearance to test image, from a large database containing fully annotated images. A structured prediction model is usually trained (e.g. CRF) to jointly model the unary potentials with plausible image models. Nearest-neighbour retrieval strategy is normally adopted in these approaches. It thus leads an intriguing trade-off between computational efficiency and accuracy. In theory, segmentation accuracy can be continuously improved by the use of ever increasing amounts of data. However, these methods have to sacrifice speed in retrieval when utilise a large database, which limits their practical applicability.

All these conventional methods tend to work well, given a sufficient amount of fully labelled data. Unfortunately, such pixel-wise labelings are very expensive and challenging to obtain.

2.2.2 Weakly Supervised Segmentation

Most existing semantic segmentation models are FS requiring pixel-level labels [118, 121, 122, 123]. A few WS semantic segmentation methods have been presented recently, exploring a variety of models such as CRF [124, 125], label propagation [126] and clustering [127]. Hanwell and Mirmehdi [23] have proposed a model for learning colour attributes from noisy, WS training data. Another closely related problem is two-class or multi-class co-segmentation [128], where the task is to segment shared objects from a set of images. Although co-segmentation does not require image labels *per-se*, they indeed assume common objects across multiple training images. Like previous semantic segmentation methods, this thesis focusses on how to learn a model to segment unseen and unlabelled test images, rather than solely segmenting the training images as in co-segmentation. Importantly, most of these previous methods only focus on object labels (nouns) or attributes (adjectives) alone. The proposed IBP based model in Chapter 4 provides a mechanism to jointly learn objects, attributes and their association (adjective-noun pairs). This thesis shows that attribute labels provide valuable complementary information via inter-label correlation, especially under this more ambitious WS setting.

It is worth noting that the proposed work is not the first one to exploit the benefit of joint modelling object and attributes for segmentation. Recently, Zheng *et al.* [123] formulated the problem of joint attribute-object image segmentation as a multi-label problem. A fully connected CRF is proposed to simultaneously segment objects and attributes. Similarly, Li *et al.* [129] proposed a novel algorithm to extract and learn attributes from segmented objects, which notably improves object classification accuracy. However, both of these methods are FS, requiring pixel-level ground truth for training. In contrast, the proposed approach can cope with weakly labelled data to alleviate the burden of strong annotation.

2.3 Attribute Learning

2.3.1 Attribute Predication and Localisation

Early work on attribute learning [22, 130] mainly focused on learning attribute predictor to describe image or objects. A supervised discriminative learning pipeline has been widely adopted to train attribute classifier in the one-vs-all fashion. These attributes are independently learned with the attribute label of the training image or objects. Learning attributes independently ignore the correlation information among them. In order to capture this cue, some recent works attempt to jointly learn all attributes [131, 132, 133, 134]. Wang *et al.* [132] and Song *et al.* [135] propose to construct a graph encoding all attributes in the attribute domain. It has been considered it as latent variables in latent SVM for categorisation. Exploiting attribute has widely proven to be effective for improving class prediction. All these works demonstrated that attribute as mid-level concepts is very useful for describing semantic entities like objects, scenes, activities, etc. Unlike these binary attributes, other work focuses on the merits of the relative attributes [136], which mine pairwise relationships between attributes of different images.

A few papers have studied how to discover local attributes [137, 138] (i.e. attribute localisation). Local attribute discovery is similar to seeking distinctive image regions. Some regions containing semantic meaning are very common and shared across categories. With local attributes, the proposed approaches can describe objects more precisely. For example, “a bird with black beak” provide more valid information than “a

black bird”. Duan *et al.* [137] propose an interactive approach to discover local attributes that are both discriminative and semantically meaningful from image datasets annotated with fine-grained category labels and object bounding boxes.

2.3.2 Attribute as Representation for Vision Tasks

Attribute learning is also the preliminary step for many other visual tasks.

image search Related to attribute-based person Re-ID is description-based person search [57, 139, 140, 141, 142]. In this case detectors for aspects of person description (e.g., clothes, soft-biometrics) are trained. Person images are then queried by description rather than by another image as in re-id. Most methods however have limited accuracy due to: (i) training on, and producing weak (image-level) annotations; and (ii) training and testing each attribute detector independently rather than jointly. For these reasons, they are also limited in being able to make complex attribute-object queries such as “Black-Jeans + Blue-Shirt”, which requires a strong joint segmentation and attribute model to disambiguate the associations between attributes.

Similar notions of attribute-based search have been exploited in a face image context [143]; and in a fashion/e-commerce context, where users search for clothing by description [52, 55, 144, 145]. Face image search however, is easier due to being more constrained and well aligned than surveillance images. Meanwhile in the fashion context, clean high resolution images are assumed; and crucially strong pixel-level annotations are typically used, which would be prohibitively costly to obtain for individual surveillance camera views. In this thesis, the proposed method bridge the gap between the clean and richly annotated fashion domain, and the noisy and scarcely annotated surveillance domain, to produce a powerful semantic representation for person re-id and search.

person Re-ID Person Re-ID is now a very well studied topic, with [50, 51] providing good reviews. Existing approaches can be broadly grouped according to two sets of criteria: supervision and representation/matching. Unsupervised approaches [146, 147, 148] are more practical in that they do not require the use of per-target camera annotation (labelling people into matching pairs), while supervised approaches [149, 150, 151, 152, 153, 154] that use this information tend to perform better. Studies also either focus on building a good representation [146, 147, 148, 149, 155, 156, 157],

or building a strong matching model [151, 152, 153, 158, 159, 160]. In this thesis, the presented approach focusses on learning mid-level semantic representations for both unsupervised and unsupervised re-id. Zhao *et al.* [161] also attempted to learn mid-level representations. However their method does not learn a *semantic* representation, so it cannot be used for description-based person search; and it relies on supervised learning. Ours can be used with or without supervision, but provides the biggest benefit in the unsupervised context. A few studies [162, 163] address person re-id based on transfer learning. However, they transfer matching models rather than representations, and between different surveillance datasets rather than different domains.

Semantic attribute-based representations have also been studied for person re-identification. A key motivation is that it is hard for low-level representations (e.g., colour/texture histograms) to provide both view-invariance and person-variance. Early studies train individual attribute detectors independently on images weakly-annotated with multi-label attributes [57, 164, 165], followed by using the estimated attribute vector of a test image as its representation for matching. More recent studies have considered joint learning of attributes and inter-attribute correlation [58, 166, 167]. Nevertheless, these approaches do not reach state-of-the-art performance. This is partly due to the drawbacks of (1) requiring ample target domain attribute annotations, which are hard to obtain at scale for learning robust detectors; and (2) producing coarse/weak (image-level) rather than strong (patch-level) annotations – due to the lack of strongly annotated training data that is even harder to obtain. In contrast, by transferring an attribute representation learned on existing large annotated fashion datasets, the proposed transfer learning based model of Chapter 5 overcomes these issues.

Attribute Learning for Describing Person Beyond person re-id and search, semantic attribute learning [4, 168] is well studied in computer vision, which is too broad to review here. However, it is worth to note that most attribute-learning studies have one or more simplifications compared to the problem this thesis considered here: They consider within-domain rather than cross-domain attributes; or produce coarse image-level attributes rather than segmenting the region of an attribute; or they do not address representing multiple attributes simultaneously on a single patch (important for e.g. a detailed clothing representation including, e.g., category + texture + colour).

Most other semantic attribute detection studies take a discriminative approach to learning for maximum detection accuracy [52, 53, 57, 139, 142, 143, 144, 169]. In con-

trast, the presented generative approach provides advantages in more naturally supporting weakly-supervised learning, and domain transfer through Bayesian priors. Related generative models include the work of Wang *et al.* [148], which used unsupervised topic models to estimate saliency for re-id, and the work of Feng *et al.* [59] which addressed data driven attribute discovery and learning.

2.4 Joint Learning of Object and Attributes.

Attributes have been used to describe objects [22, 170], people [171], clothing [144], scenes [138], faces [6], and video events [169]. However, most previous studies learn and infer object and attribute models separately, e.g., by independently training binary classifiers, and require strong annotations/labels indicating object/attribute locations and/or associations if the image is not dominated by a single object. A few recent studies have learned object-attribute association explicitly [1, 131, 132, 138, 172, 173, 174]. Different from the proposed approach, [131, 132, 172, 174] only train and test on unambiguous data, i.e. images containing a single dominant object, assuming object-attribute association is known at training; moreover, they allocate exactly one attribute per object. Kulkarni *et al.* [1] model the more challenging PASCAL VOC type of data with multiple objects and attributes coexisting. However, their model is pre-trained on object and attribute detectors learned using strongly annotated images with object bounding boxes provided. The work in [138] also does object segmentation and object-attribute prediction. But its model is learned from strongly labelled images in that object-attribute associations are given during training; and importantly prediction is restricted to object-attribute pairs seen during training. In summary none of the existing work learns object-attribute association from weakly labelled data as the proposed approach does in this thesis.

Some existing works aim to perform attribute-based query [6, 27, 175, 176]. In particular, recent studies have considered how to calibrate [176] and fuse [27] multiple attribute scores in a single query. This thesis goes beyond these studies in supporting conjunction of object+multi-attribute query. Moreover, existing methods either require bounding boxes or assume simple data with single dominant objects, and do not reason jointly about multiple attribute-object association. This means that they would be intrinsically challenged in reasoning about (multi)-attribute-object queries on chal-

lenging data with multiple objects and multiple attributes in each image (e.g., querying furry brown horse, in a dataset with black horses and furry dogs in the same image). In other words, they cannot be directly extended to solve query by object-attribute association.

2.5 Related Work in Weakly Supervised Learning

The works discussed in previous sections are reviewed according to the related computer vision tasks. In this section, the thesis provides a detailed comparison between the proposed approach and related WSL based approaches.

2.5.1 Discriminative Models

Discriminative methods underpin many high performance recognition and annotation studies [1, 22, 138, 171, 177, 178]. Similarly existing WSL methods are also dominated by discriminative models. Apart from the above mentioned CRF [124, 125], label propagation [126] and clustering [127] models, some discriminative MIL models were also adopted [42, 96]. the designed model is a probabilistic generative model. Compared to a discriminative model, the main strength of a generative model under the WS setting is that it is able to infer latent factors corresponding to background clutter and un-annotated object/attributes and modelling them jointly in a single model so as to explain away the ambiguity existing in the weak image-level labels. Very recently deep learning based automated image captioning has started to attract attention [5, 26]. Generating a natural sentence to describing the content of image obviously is a harder task as apart from nouns and adjective, other words including verb (action) and preposition (where) need to inferred and language syntax needs to be followed in the generated text description. However, these neural network models are essentially still discriminative models and have the same drawbacks as other discriminative models for WSL.

2.5.2 Probabilistic Generative Models

LTM were originally developed for unsupervised text analysis [179], and have been successfully adapted to both unsupervised [79, 180] and supervised image understanding problems [181, 182, 183, 184, 185]. Most studies have addressed the simpler tasks of classification [184, 185] and annotation [184, 186]. The proposed joint localisation model of Chapter 3 differs from the existing ones in two main aspects: (i) Conventional topic models have no explicit notion of the spatial location and extent of an object in an image. This is addressed in the proposed model by modelling the spatial distribution of each topic. Note that some topic model based methods [182, 183] can also be applied to WSOL. However, the spatial location is obtained from a pre-segmentation step rather than being explicitly modelled. (ii) The other difference is more subtle – existing supervised topic models such as CorrLDA [186], SLDA [184] and derivatives [183] only weakly influence the learned topics. This is because the objective is the sum of visual words and label likelihoods, and visual words vastly outnumber annotations, thus dominating the result [185]. The limitation is serious for WSOL as the labels are already weak and they must be used to their full strength. In this thesis, a learning algorithm with topic constraints similarly to [29] is formulated to provide stronger supervision which is demonstrated to be much more effective than the conventional supervised topic models in various experiments (see Appendix A). With these limitations addressed, the presented method can exploit the potential of a generative model for domain adaptation, joint-learning of multiple objects and SSL.

The flexibility of generative probabilistic models and their suitability particularly for WSL has seen them successfully applied to a variety of WSL tasks [169, 183, 187, 188]. These studies often generalise Probabilistic Topic Model (PTM) [179]. However PTMs are limited for explaining objects and attributes in that latent topics are competitive - the fundamental assumption is that an object is a horse *or* brown *or* furry. They intrinsically do not account for the reality that it is *all* at once. In contrast, The proposed model in Chapter 4 generalises IBP [49, 189]. The IBP is a latent feature model that can independently activate each latent factor, explaining imagery as a weighted sum of active factor appearances.

As mentioned earlier, the presented WS-MRF-SIBP differs significantly from the standard flat and unsupervised IBP in that it is hierarchical to model grouped data (im-

ages composed of superpixels) and WS. This allows us to exploit image-level weak supervision, but disambiguate it to determine the best explanation in terms of which superpixels correspond to un-annotated background, which superpixels corresponds to which annotated objects, and which objects has which attributes. In addition, a MRF is integrated into the IBP to model correlations of factors both within and across superpixels. A few previous studies [190, 191, 192] generalise the classic Latent Dirichlet Allocation (LDA) topic model [179] by integrating a MRF to enforce the spatial coherence across topic labels of neighbouring regions. Unlike these methods, the proposed approach generalises the IBP by defining the MRF over hidden factors. Furthermore, beyond encoding spatial coherence the designed method also defines a factorial MRF to capture attribute-attribute and attribute-object co-occurrences within superpixels.

An approach similar in spirit to ours in the sense of jointly learning a model for all classes is that of Cabral *et al.* [193]. This study formulates multi-label image classification as a matrix completion problem, which is also related to the presented factoring images into a mixture of topics. However the proposed joint localisation model of Chapter 3 add two key components of (i) a stronger notion of the spatial location and extent of each object, and (ii) the ability to encode human knowledge or transferred knowledge through a Bayesian prior. As a result, the proposed model is able to address more challenging data than [193] such PASCAL VOC. MIML [96] approaches provide a mechanism to jointly learn a model for all classes [194, 195]. However, because these methods must search for a discrete space (of positive instance subsets), their optimisation problem is harder than the smooth probabilistic optimisation here. Finally, while more elaborate joint generative learning methods [183, 196] exist, they are more complicated than necessary for WSOL and do not scale to the size of data required here.

2.5.3 Exploiting Prior Knowledge

Prior knowledge has been exploited in existing WSOL works [42, 43, 44]. Recognition or detection priors can be broadly broken down into appearance and geometry (location, size, aspect) cues, and can be provided manually, or estimated from data. The most common use has been crude: to generate candidate object locations based on a pre-trained model for generic objectness [98], i.e. the previously mentioned saliency

cue. This reduces the search space for discriminative models. Beyond this, geometry priors have also been estimated during learning [42]. The proposed joint localisation model of Chapter 3 not only can exploit such simple appearance and geometry cues as model priors, but also go beyond to exploit a richer object hierarchy, which has been widely exploited in classification [47, 197, 198, 199] and to a less extent detection [34, 38]. More specifically, the proposed joint localisation model of Chapter 3 leverages WordNet, a large lexical database based on linguistics [200]. WordNet provides a measure of prior appearance similarity/correlation between classes, and the proposed method uses this prior to regularise appearance learning. Such cross-class appearance correlation information is harder to use in previous WSOL approaches because different classes are trained separately. Interestingly, the proposed model uniquely shows positive results for WordNet-based appearance correlation, in contrast to some recent studies [47, 199] that found no or limited benefit from exploiting WordNet based cross-class appearance correlation for recognition. Compared to the classification task, this inter-class correlation information is more valuable for WSOL because the task is more ambiguous. Specifically, the interdependent localisation and appearance learning aspects of the task adds an extra layer of ambiguity – the model might be able to learn the appearance if it knew the location, but it will never find the location without knowing appearance. The proposed joint localisation model of Chapter 3 is also related to [38] where hierarchical cross-class appearance similarity is used to help WSOL in ImageNet by transfer learning. However, a source dataset of fully annotated images are required in their work, whilst the designed model exploits the correlation directly for the target data which is only weakly labelled.

2.5.4 Cross Domain Learning

Domain adaptation [201] methods aim to exploit prior knowledge from a source domain/dataset to improve the performance and/or reduce the amount of annotation required in a target domain/dataset (see [202] for a review). Many conventional approaches are based on SVMs for which the target domain can be considered a perturbed version of the source domain, and thus learning proceeds in the target domain by regularising it toward the source [39]. More recently, transductive SVM [203], Multiple Kernel Learning (MKL) [204], and instance constraints [205] have been exploited. In

contrast to these discriminative approaches, the proposed joint localisation model of Chapter 3 exploits a simple and efficient Bayesian adaptation approach similar in spirit to [201, 206]. Posterior parameters from the source domain are transferred as priors for the target, which are then adapted based on observed target domain data via Bayesian learning. Going beyond simple within-modality dataset bias, recent studies [40, 97] have adapted object detectors from video to image or reverse. This thesis shows that the proposed joint localisation model of Chapter 3 can achieve the image-video domain transfer within a single framework.

A key challenge for the proposed transfer learning based model of Chapter 5 is the domain-shift between the fashion and surveillance domains. Addressing a change of domains with domain adaptation is well studied in computer vision [207, 208, 209] and beyond [202]. To avoid necessitating target domain annotation, this task requires unsupervised adaptation which is harder. Some off-the-shelf solutions exist [207, 210], but these under-perform due to working blindly in the low-level feature space, disconnected to the semantics being modelled, i.e. attributes. In contrast, in the style of [201, 209, 211], the proposed method achieves domain adaptation by transferring the source domain attribute model as a prior when the target domain model is learned. This enables adaptation to the target domain, while exploiting the constraints provided by semantic attribute model.

2.5.5 Exploiting Unlabelled Data

SSL [48] methods aim to reduce labelling requirements and/or improve results compared to only using labelled data. Most existing SSL approaches assume a training set with a mix of fully labelled and weak or unlabelled [38, 212] data, while the presented method uses weak and unlabelled data alone. The existing (discriminative) line of work focusing on WSOL [42, 44, 213, 214] has not generally exploited unlabelled data, and cannot straightforwardly do so.

2.5.6 Feature Fusion

Combining multiple complementary cues has been shown to improve classification performance in object recognition [204, 215, 216, 217]. Two simple feature fusion

methods have been widely used in existing work: early fusion which combines low-level features [218] early (feature concatenation) and late (score level) fusion [42, 90]. MKL approaches have attracted attention as a principled mid-level approach to combining features [216, 217]. Similarly to MKL, the proposed joint localisation model of Chapter 3 provides a principled and jointly-learned mid-level probabilistic fusion via its generative process.

2.6 Summary

A FSL strategy is typically required by most existing approaches for object recognition, attribute learning and other image understanding related tasks. The detailed annotation reduces the visual ambiguity by disentangles object instances from a noisy and cluttered background. Nevertheless, it puts a heavy burden on the user to provide sufficient examples of the object of interest, with manual annotation of object location in all images. Recently there has been an increasing interest in WS approaches to learning object appearance which does not require manual annotation of object locations. WSL provides a more scalable solution for understanding images with the prevalence of media sharing websites, such as Flickr, where a large number of social images with user provided image-level labels are available from the Internet.

Existing WSL methods have shown promising results in various image understanding tasks. Nevertheless, there are several open problems and limitations that need to be solved. Firstly, most existing WSL approaches are dominated by discriminative models. These algorithms are inspired by the success of discriminative methods in many FS computer vision tasks. These methods are hard to integrate prior knowledge and multiple related cues. This is particularly important in WSL setting where very limited information is available. Secondly, most previous WS tend to model each category independently (e.g. object detector and attribute classifier have been trained separately for every class). In fact, object categories in real work do not follow a flat structure. A more complex and hierarchical relation exists in a large number of object categories. For example, the class “motorbike” is more similar to “bike” than “boat”, meanwhile “motorbike” and “bike” belong to the “wheeled vehicle”. Learning all these categories independently ignore this rich information embedded in the data. A joint learning framework enables us to sufficiently leverage various cues and contexts including

feature-label association, inter-label correlation, spatial neighbourhood cues, and label consistency. Thirdly, there are some generative based studies devoted to jointly learning all categories. However, explicitly modelling and explaining complex semantics is still non-trivial and has remained unexplored. It is extremely challenging to understand images for each object instance from the image-level weak label due to the label ambiguity. One instance/object can be described by many (potentially unlimited number of) attributes concurrently (e.g. a car could be “red”, “metal” and “wheeled”, as well as “shiny”). It is the key to systematically and effectively incorporate all types of relation and cues for achieving a significant improvement.

In the subsequent chapters of this thesis, a number of models are formulated to address these limitations following the three key concepts below:

1. *Localising object jointly* - Existing WS object detection approaches are dominated by discriminative model, where a decision boundary is learned independently for each object category. The proposed joint localisation model of Chapter 3 jointly models all object categories together to leverage various information cues and meanwhile the discriminative power is still maintained to outperform previous approaches.
2. *Learning object and attribute jointly* - Learning object and attribute jointly has been attempted by previous studies. Most of them rely on FS training data to learn object-attribute correlation explicitly. Few prior works aim to model object and attributes from weakly labelled data. However, complex visual semantics in images make them ill-suited for holistically capturing and retaining all information. The proposed IBP based model in Chapter 4 jointly learn multiple objects, attributes and background clutter in a single framework and ambiguity in each is explained away by knowledge of the other.
3. *Transferring object and attribute jointly* - WSL can be tackled in a transfer learning angle. There are always some fully labelled or weakly labelled data available in a related domain when insufficient labelled data from the target domain. Unlike previous approach, which learn each factor (objects, attributes, other annotations) individually and adapted to target domain data, the proposed transfer learning based model of Chapter 5 further provide a solution to transfer a power-

2. Literature Review

ful semantic description including object, attribute and other automatically discovered latent factors for various computer vision tasks on weakly labelled data.

Chapter 3

Bayesian Joint Modelling for Object Localisation

3.1 Overview

Object detection is the first and essential step to understanding the content of image. This chapter considers the problem of detecting and localising objects of a generic category, such as people or cars, in an static images or video. Large scale object recognition has received increasing interest in the past five years [28, 29, 30]. Due to the prevalence of online media sharing websites such as Flickr, a lack of images for learning is no longer the barrier. A new bottleneck appears instead: the lack of annotated images, particularly strongly annotated ones. For example, for many vision tasks such as object classification [32], detection [16], and segmentation [30, 34] hundreds or even thousands of object samples must be annotated from images for each object class. This annotation includes both the presence of objects and their locations, typically in the form of bounding boxes. This is a tedious and time-consuming process that prevents tasks such as object detection from scaling to thousands of classes [38].

To this end, this chapter proposes a WSOL approach, which simultaneously locates objects in images and learns their appearance using weak labels indicating only the presence/absence of the object of interest. Specifically, both multiple object classes and different types of backgrounds are modelled jointly in a single generative model as latent topics, in order to explicitly exploit their correlations (see Figure 3.1). As

BoW models, conventional LTMs have no notion of localisation. The proposed model overcomes this problem by incorporating an explicit notion of object location, alongside the ability to incorporate prior knowledge about the object appearance in a fully Bayesian approach. Importantly, as a joint generative model, unlabelled data can now be easily used to compensate for sparse training annotations, simply by allowing the model to also infer both which unknown objects are present in those images and where they are. WSOL approaches first locate the object of interest in the training images and then the location information is used to train a detector in a FS fashion.

This chapter is structured as follows: the proposed framework is explained in Section 3.2. Section 3.4 described the core inference process including label-topic constraints and probabilistic fusion. Section 3.5 demonstrates that how the learned model can be applied to localise object in images and videos. The various prior knowledge that embedded in the proposed model are discussed in Section 3.6. In Section 3.7, a solution of utilising additional data is further provided to transfer the knowledge from the other domain. Extensive results are reported and discussed in Section 3.8. Finally, conclusions are drawn in Section 3.9.

3.2 Joint Topic Model for Objects and Background

In this section, a new LTM [179] approach is introduced to the WSOL task. Applied to images, conventional LTMs factor images into combinations of latent topics [79, 180]. Without supervision, these topics may or may not correspond to anything of semantic relevance to humans. To address the WSOL task, the ideal model needs to learn what is unique to all images sharing a particular label (object class), while explaining away both the pixels corresponding to other annotated objects as well as other shared visual aspects (background) which are irrelevant to the annotations of interest. The presented method will achieve this in a LTM framework by applying weak supervision to partially constrain the available topics for each image. This constraint is enforced by label/topic clamping to ensure that each foreground topic corresponds to an object class of interest.

More specifically, to address the WSOL task, the proposed approach will factor images into unique combinations of K shared topics. If there are C classes of objects to be localised, $K^{fg}=C$ of these will represent the (foreground) classes, and

3. Bayesian Joint Modelling for Object Localisation

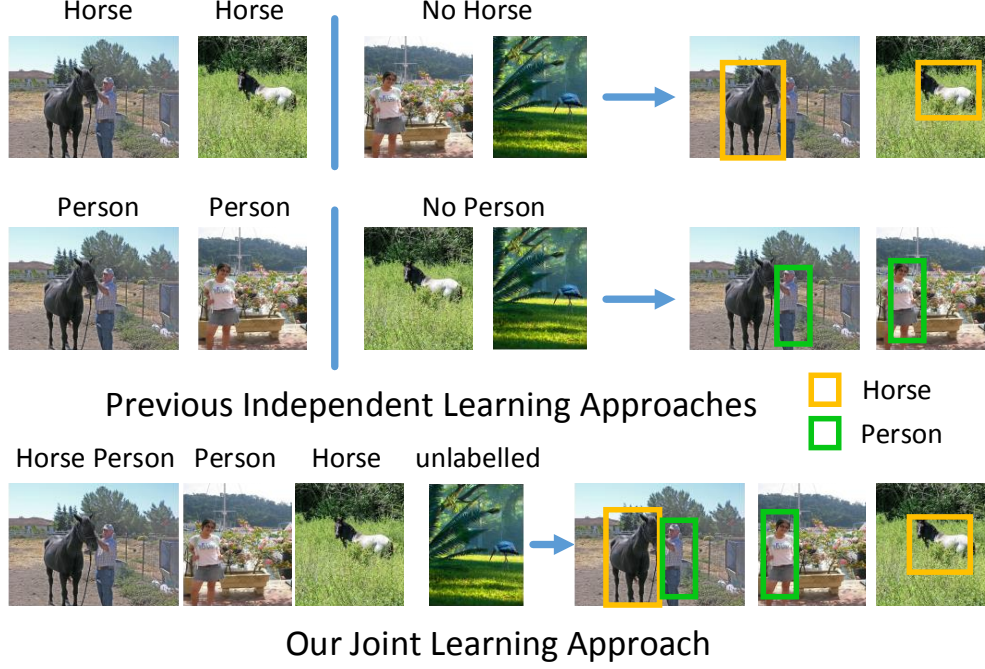


Figure 3.1: Different types of objects often co-exist in a single image. The proposed joint learning approach differs from previous approaches which localise each object class independently.

$K^{bg} = K - K^{fg}$ topics will model background data to be explained away. Each topic thus corresponds to one object class or one type of background. Let T^{fg} and T^{bg} index foreground and background topics respectively. An image is represented using a Bag-of-Words (BoW) representation for each of $f = 1 \dots F$ different types of features (see Section 3.8.1 for the specific appearance features used). After learning, each latent topic will encode both a distribution over the V_f sized appearance vocabulary of each feature f and also over the spatial location of these words within each image. Formally, given a set of J training images, each labelled with any number of the C foreground classes, and represented as bags of words \mathbf{x}_{jf} , the generative process of the proposed model (Figure 3.3) is as follows :

For each topic $k \in 1 \dots K$:

1. For each feature representation $f \in 1 \dots F$:

3. Bayesian Joint Modelling for Object Localisation

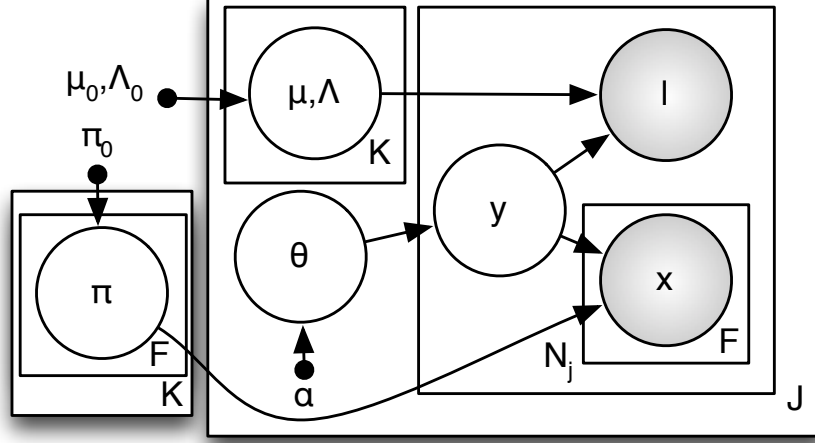


Figure 3.2: Graphical model for the proposed WSOL joint topic model. Shaded nodes are observed.

- (a) Draw an appearance distribution $\pi_{kf} \sim \text{Dir}(\pi_{kf}^0)$ following the Dirichlet distribution

For each image $j \in 1 \dots J$:

1. Draw foreground and background topic distribution $\theta_j \sim \text{Dir}(\alpha_j)$, $\alpha_j = [\alpha_j^{fg}, \alpha_j^{bg}]$ where the Dirichlet distribution parameter α_j reflects prior knowledge of the presence of each object class or background in the image j . Both θ_j and α_j are K dimensional.
2. For each foreground topic $k \in T^{fg}$ draw a location distribution:

$$\{\mu_{kj}, \Lambda_{kj}\} \sim \mathcal{NW}(\mu_k^0, \Lambda_k^0, W_k^0, \xi_k^0)$$
3. For each observation (visual word) $i \in 1 \dots N_j$:
 - (a) Draw topic $y_{ij} \sim \text{Multi}(\theta_j)$
 - (b) Draw a location:
$$l_{ij} \sim \mathcal{N}(\mu_{y_{ij}j}, \Lambda_{y_{ij}j}^{-1}) \text{ if } y_{ij} \in T^{fg} \text{ or}$$

$$l_{ij} \sim \text{Uniform} \text{ if } y_{ij} \in T^{bg}$$
 - (c) For each feature representation $f \in 1 \dots F$:

3. Bayesian Joint Modelling for Object Localisation

i. Draw visual word $x_{ijf} \sim \text{Multi}(\pi_{y_{ijf}})$

where Multi , Dir , \mathcal{N} , \mathcal{NW} and Uniform respectively indicate Multinomial, Dirichlet, Normal, Normal-Wishart and uniform distributions with the specified parameters. These prior distributions are chosen mainly because they are conjugate to the word, topic and location distributions, and hence enable efficient inference. For the visual word spatial location, the foreground and background distributions are of different forms – normal for foreground and uniform for background. This is to reflect the intuition that foreground objects tend to be compact and background much less so. The joint distribution of all observed $O = \{\mathbf{x}_{jf}, \mathbf{l}_j\}_{j,f=1}^{J,F}$ and latent $H = \{\{\pi_{kf}\}_{k,f=1}^{K,F}, \{\mathbf{y}_j, \boldsymbol{\mu}_{kj}, \Lambda_{kj}, \boldsymbol{\theta}_j\}_{k,j=1}^{K,J}\}$ variables given parameters $\Pi = \{\{\pi_{kf}^0\}_{k,f=1}^{K,F}, \{\boldsymbol{\mu}_k^0, \Lambda_k^0, \mathbf{W}_k^0, \xi_k^0\}_{k=1}^K, \{\boldsymbol{\alpha}_j\}_{j=1}^J\}$ in the proposed model is therefore:

$$p(O, H|\Pi) = \prod_k^K \prod_f^F p(\pi_{kf} | \pi_{kf}^0) \quad (3.1)$$

$$\cdot \prod_j^J p(\boldsymbol{\theta}_j | \boldsymbol{\alpha}_j) \left[\prod_k^K p(\boldsymbol{\mu}_{jk}, \Lambda_{jk} | \boldsymbol{\mu}_k^0, \Lambda_k^0, \mathbf{W}_k^0, \xi_k^0) \left(\prod_i^{N_j} p(\mathbf{l}_{ij} | \boldsymbol{\mu}_{jk}, \Lambda_{jk}^{-1}) \prod_f^F p(x_{ijf} | y_{ij}, \pi_{y_{ijf}}) p(y_{ij} | \boldsymbol{\theta}_j) \right) \right]. \quad (3.2)$$

3.3 Relationship with Latent Dirichlet Allocation

In this section we compare the proposed model with LDA [179] to highlight the key differences and similarities. The LDA model is represented as a probabilistic graphical model in Figure . Similar to ours, there are three levels to the LDA representation. In order to address the WSOL problem, this work introduces the following features/improvements in terms of both model structure and learning algorithm:

1. **Effective supervision:** The problem of how to achieve scalable and accurate supervision in topic models is addressed. Unlike the unsupervised LDA that adopts the empirical Bayes approach to estimating parameter α , the proposed model use α to encode the supervision from weak labels.

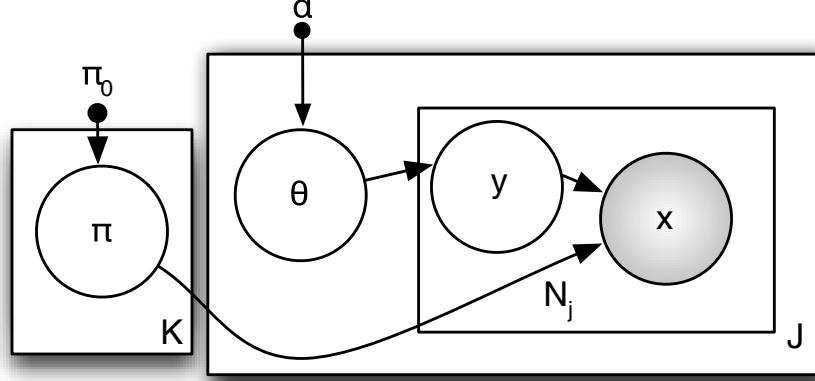


Figure 3.3: Graphical model for the classic latent dirichlet allocation. Shaded nodes are observed.

2. **Two sets of topics:** this work learned with different levels of supervision - one sets are weakly supervised by the weak labels, and the other discovered unsupervised. Used together in this way they can jointly learn foreground and background.
3. **Initialisation of model priors:** effective establishment of informative priors (π_{kf}^0) is important especially for this weakly supervised problem (multiple objects co-existing with cluttered background), where location ambiguity would otherwise make it very easy for the model to converge to inappropriate local minima in learning.
4. **Spatial object location:** ($\mu_{kj}, \Lambda_{kj}^{-1}$) represents object location explicitly (which is not natural in topic models), which enables objects to be well learned even when their appearance is non-uniform or they are of ambiguous appearance.

3.4 Model Learning

Inference via variational message passing Learning the presented model involves inferring the following quantities: the appearance of each object class for each fea-

3. Bayesian Joint Modelling for Object Localisation

ture type, $\pi_{kf}, k \in T^{fg}$ and each background type, $\pi_{kf}, k \in T^{bg}$ for each feature type f ; the word-topic distribution (soft segmentation) of each image y_j , the proportion of visual words (related to the proportion of pixels) in each image corresponding to each class or background θ_j , and the location of each object μ_{jk}, Λ_{jk} in each image (mean and covariance of a Gaussian). To learn the model and localise all the weakly annotated objects, the model wish to infer the posterior $p(H|O, \Pi) = p(\{\mathbf{y}_j, \mu_{jk}, \Lambda_{jk}, \theta_j\}_{k,j}^{K,J}, \{\pi_{kf}\}_{k,f}^{K,F} | \{\mathbf{x}_{jf}, \mathbf{l}_j\}_{j=1, f=1}^{J,F}, \Pi)$. This is intractable to solve directly; however a Variational Message Passing (VMP) [219] strategy can be used to obtain a factored approximation $q(H|O, \Pi)$ to the posterior:

$$q(H|O, \Pi) = \prod_{k,f} q(\pi_{kf}) \prod_j q(\theta_j) q(\mu_{jk}, \Lambda_{jk}) \prod_i q(y_{ij}). \quad (3.3)$$

Under this approximation a VMP solution is obtained by deriving integrals of the form $\ln q(\mathbf{h}) = E_{H \setminus \mathbf{h}} [\ln p(H, O)] + K$ for each group of hidden variables \mathbf{h} , thus obtaining the following updates for the sufficient statistics (indicated by tilde) of each variable:

$$\tilde{\theta}_{jk} = \alpha_{jk} + \sum_i \tilde{y}_{ijk}, \quad (3.4)$$

$$\begin{aligned} \tilde{y}_{ijk} &\propto \int_{\mu_{jk}, \Lambda_{jk}} \mathcal{N}(\mathbf{l}_{ij} | \mu_{jk}, \Lambda_{jk}^{-1}) q(\mu_{jk}, \Lambda_{jk}) \\ &\cdot \prod_f \exp \left(\Psi(\tilde{\pi}_{x_{ijf} y_{ijf}}) - \Psi(\sum_v \tilde{\pi}_{v y_{ijf}}) \right) \\ &\cdot \exp \left(\Psi(\tilde{\theta}_{j y_{ijk}}) \right), \end{aligned} \quad (3.5)$$

$$\tilde{\pi}_{v kf} = \pi_{v kf}^0 + \sum_{ij} \mathbf{I}(x_{ijf} = v) \tilde{y}_{ijk}, \quad (3.6)$$

where Ψ is the digamma function, $v = 1 \dots V_f$ ranges over the BoW appearance vocabulary, \mathbf{I} is the indicator function which returns 1 if its argument is true, and the integral in second line returns the student-t distribution over \mathbf{l}_{ij} , $\mathcal{S}(\mathbf{l}_{ij} | \tilde{\mu}_{jk}, \tilde{\Lambda}_{jk}^{-1}, \tilde{\mathbf{W}}_{jk}, \tilde{\xi}_{jk})$. Within each image j , standard updates [220] apply for the sufficient statistics $\{\tilde{\mu}_{jk}, \tilde{\Lambda}_{jk},$

3. Bayesian Joint Modelling for Object Localisation

$\tilde{W}_{jk}, \tilde{\xi}_{jk}$ of Normal-Wishart parameter posterior $q(\mu_{jk}, \Lambda_{jk})$. The update in Eq. (3.5) (estimating the object explaining each pixel) is the most non-standard for LTMs; this is because it contains a top-down contribution (the third term), and two bottom-up contributions from the location and appearance (the first and second terms respectively). The model is learned by iterating the updates of Eqs. (3.4)-(3.6) for all images j , words i , topics k and vocabulary v .

Supervision via label-topic constraints In conventional topic models, the α parameter encodes the expected proportion of words for each topic. In the proposed WS topic model, the designed model uses α to encode the supervision from weak labels. In particular, α_j^{fg} is set to a binary vector with $\alpha_{jk}^{fg} = 1$ if class k is present in image j and $\alpha_{jk}^{fg} = 0$ otherwise. α^{bg} is always set to 1 to reflect the fact that background of different types can be shared across different images. That is, the foreground topics are clamped with the weak labels indicating the presence/absence of foreground object classes in each image, whilst all background types are assumed to be present a priori. With these partial constraints, iterating the updates in Eqs. (3.4)-(3.6) has the effect of factoring images into combinations of latent topics, where K^{bg} background topics are always available to explain away backgrounds, and K^{fg} foreground topics are only available to images with annotated classes. Note that this set-up assumes a 1:1 correspondence between object classes and topics. More topics can trivially be assigned to each object class (1:N correspondence), which has the effect of modelling multi-modality in object appearance, for a linear increase in computational cost.

Probabilistic feature fusion The proposed method combines multiple features probabilistically. A single topic distribution (y) is estimated given different low-level features (f) in Eq. (3.5). the proposed fusion strategy keeps the original low-level feature representations rather than increasing ambiguity by concatenating them before they provide complementary information about the location (early fusion). The shared topic (y) and Gaussian location distribution (μ, Λ^{-1}) correlate the multiple features, which avoids domination by a single one. The appearance model in each modality is updated based on the consensus estimate of location; it thus learns a good appearance in each view even if the particular category is hard to detect in that view (as a result could drift if used alone). Its advantage over early (feature concatenation) or late (score level) fusion is demonstrated experimentally in Section A.1 of the Appendix A.

3.5 Object Localisation

After learning, the designed method extracts the location of the objects in each image from the model, which can then be used to learn an object detector. Depending on whether the images are treated as individual images or consecutive video frames, the proposed localisation method differs slightly.

Individual images There are two possible strategies to localise objects in individual images, which will be compared later in Section 3.8. In the first strategy (*Our-Gaussian*), a bounding box for class k in image j can be obtained directly from the Gaussian mean of the parameter posterior $q(\boldsymbol{\mu}_{jk}, \Lambda_{jk})$, via aligning a bounding box to the two standard deviation ellipse. This has the advantage of being clean and highly efficient. However, since there is only one Gaussian per class (which will grow to cover all instances of the class in an image), this is not ideal for images with more than one object per class. In the second strategy (*Our-Sampling*) a heat-map is drawn for class k by projecting $q(y_{ijk})$ (Eq. (3.5)) back onto the image plane, using the grid coordinates where visual words are computed. This heat-map is analogous to those produced by many other approaches such as Hough transforms [221]. Thereafter, any strategy for heat-map based localisation may be used. A Non-maximum Suppression (NMS) strategy of [16] is adopted here.

Video frames Here we provide two ways to extend the proposed method on video data.

1. One video is composed of multiple frames. When we treat each frame separately, this is exactly the same form as individual images. Thus, the two strategies used on individual images are directly applicable to video data.
2. However, the temporal information of objects is useful in continuous videos to smooth the noise of individual frames. To this end, this model apply a simple state space model for video segments to post-process object locations, smoothing them in time. Two diagonal points are sufficient to encode object location (bounding-box), and these are estimated from $q(\boldsymbol{\mu}_{jk}, \Lambda_{jk})$ above at every frame/time t as c_t . Assuming a four-dimensional state latent state vector $\mathbf{z}_t^T = (z_{xt} \ z_{yt} \ \dot{z}_{xt} \ \dot{z}_{yt})$, denoting the (hidden) true coordinate of an object of interest (two diagonal corners of the bounding box). A Kalman smoother is then adopted

3. Bayesian Joint Modelling for Object Localisation

to smooth the observation noise σ_t in the system:

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \epsilon_t, \mathbf{c}_t = \mathbf{O}\mathbf{z}_t + \sigma_t, \quad (3.7)$$

where \mathbf{A} is the temporal transition between true locations \mathbf{z} in video, and \mathbf{O} is the observation function for each frame.

3.6 Bayesian Priors

An important capability of the proposed Bayesian approach is that top-down cues from human expertise, or estimated from data can be encoded. Various types of human knowledge about objects and their relationships with background are encoded in the proposed model. As discussed earlier, prior cues can potentially cover appearance and geometry information.

Encoding geometry prior For geometry, the proposed approach already models the most general intuition that objects are compact relative to background by assigning them Gaussian and uniform distributions respectively (Section 3.2). Beyond this, prior knowledge about typical image location and size of each class can be included via prior parameters μ_k^0, Λ_k^0 , however this thesis found this did not actually noticeably improve results in the presented experiments so the proposed method did not exploit it. This makes sense, because in challenging datasets like PASCAL VOC, objects appear in highly variable scales and locations, so there is little regularity to learn.

Encoding appearance prior If prior information is available about object category appearance, it can be included by setting π_{kf}^0 . (The presented method will exploit this later for cross-domain adaptation in Section 3.7.1). For within-domain learning, the proposed approach can obtain an initial data-driven estimate of object appearance to use as a prior by exploiting the observation that, when aggregated across all images, the background is more dominant than any single object class in terms of size (hence the amount of visual words). Exploiting this intuition, for each object class k , the

3. Bayesian Joint Modelling for Object Localisation

appearance prior π_{kf}^0 is set as:

$$\pi_{kf}^0 = \left| \frac{1}{C} \sum_{j, c_j=k} h(\mathbf{x}_{jf}) - \frac{1}{J} \sum_j h(\mathbf{x}_{jf}) \right|_+ + \epsilon, \quad (3.8)$$

where $h(\cdot)$ indicates histogram and ϵ is a small constant. That is, set the appearance prior for each class to the mean of those images containing the object class minus the average over all images. This results in a prior which reflects what is consistently unique about each particular class. This is related to the notion of saliency, not within an image, but across all images. Saliency has been exploited in previous MIL based approaches to generate the instances/candidate object locations [42, 43, 44, 90, 91]. However, in the presented model it is cleanly integrated as a prior.

Encoding appearance similarity prior Going beyond the direct unary appearance prior discussed above, the presented method next considers exploiting the notion of prior *inter-class appearance similarity*, rather than prior appearance per-se. The prior similarity between each object category can be estimated by computing inter-category category distance based on WordNet structure [200]. The proposed model computes a similarity matrix \mathcal{M} where elements $\mathcal{M}_{m,n}$ indicates the relatedness between class m and n . The similarity matrix is then used to define how much appearance information from class m contributes to class n a priori.

The presented method exploits this matrix by introducing an M-step into the proposed learning algorithm process (Eqs. (3.4)-(3.6)). Previously the appearance prior π_{kf}^0 was considered fixed (e.g., from Eq. (3.8)). As with any parameter learning in the presence of latent variables, π_{kf}^0 could potentially be optimised by a maximum-likelihood M-step interleaved with E-step latent variable inference. However, rather than the conventional approach of optimising π_{kf}^0 *solely* given the data of class k , the presented approach defines an update that exploits cross-class similarity by updating π_{kf}^0 using *all* the data, but weighted by its similarity to the target class k .

Denoting $\hat{\pi}_{vkf}^0$ as the new appearance prior to be learned, the proposed method introduces a new regularised M-step to learn $\hat{\pi}_{vkf}^0$. Specifically, the update for each

class $k \in T^{fg}$ is as follows:

$$\hat{\pi}_{vkf}^0 = \underbrace{\pi_{vkf}^0}_{\text{fixed data driven prior}} + \underbrace{\sum_{ij} \sum_{k' \in T^{fg}} \mathcal{M}_{k,k'} \cdot \mathbf{I}(x_{ijf} = v) \tilde{y}_{ijk'}}_{\text{inter-class similarity prior}} \quad (3.9)$$

The first term π_{vkf}^0 is the original unary prior from Eq. (3.8). The second term is a data-driven update given the results of the E-step (\tilde{y} , Eqs. (3.4)-(3.6)). It includes a contribution from all images of all classes k' , weighted by the similarity of k' to the target class k – given by $\mathcal{M}_{k,k'}$. The updated $\hat{\pi}_{kf}^0$ then replaces π_{kf}^0 in Eq. (3.6) of the E-step.

3.7 Learning from Additional Data

This section discusses learning from additional data beyond the data for the WSOL task. This includes partially relevant data from other domains or datasets, and any additional but un-annotated data from the same domain.

3.7.1 Bayesian Domain Adaptation

Across different datasets or domains (such as images and video), the appearance of each object category will exhibit similarity, but vary sufficiently that directly using an appearance model learned in a source domain s for inference in a target domain t will perform poorly [41]. In this case it would correspond to directly applying a learned source appearance model π_k^s to a new target domain t , $\pi_k^t := \pi_k^s$. However, one hopes to be able to exploit similarities between the domains to learn a better model than using only the target domain alone [39, 97, 203, 206]. In this case, the Bayesian (Multinomial-Dirichlet conjugate) form of the proposed model is able to achieve this for WSOL by simply learning π_k^s for a source domain s (Eq. (3.6)), and applying it as the prior $\pi_k^{0t} := \pi_k^s$ in the target t – which is then adapted to reflect the target domain statistics (Eq. (3.6)).

3.7.2 Semi-supervised Learning

Beyond learning from annotated data in different but related domains, the proposed framework can also be applied in a SSL context to learn from unlabelled data in the same domain to improve performance and/or reduce annotation requirement. Specifically, images j with known annotations are encoded as described in Section 3.4, while those without annotation are set to $\alpha_j^{fg} = 0.1 \forall j$, meaning that all topics/classes may potentially occur, but it expects few simultaneously within one image. Unknown images can include those from the same pool of classes but without annotation (for which the posterior $q(\theta)$ will pick out the present classes), or those from a completely disjoint pool of classes (for which $q(\theta)$ will encode only background).

3.8 Experiments

3.8.1 Datasets, Features and Settings

Datasets This section evaluates the proposed model on three datasets, PASCAL VOC [18], ImageNet [19] and YTO video [40]. The challenging PASCAL VOC 2007 dataset is now widely used for WSOL. A number of variants are used: *VOC07-20* contains all 20 classes from VOC 2007 training set as defined in [90] and was used in [43, 90, 91]; *VOC07-6×2* contains 6 classes with Left and Right poses considered as separate giving 12 classes in total and was used in [42, 43, 44, 46, 90, 91]. The former obviously is more challenging than the latter. Note that *VOC07-20* is different to the *Pascal07-all* defined in [42] which actually contains 14 classes and uses the other 6 as fully annotated auxiliary data. It is called *VOC07-14* for consistency, but does not use the other 6 auxiliary classes.

To evaluate the proposed method in a larger-scale setting, this thesis selects all images with bounding box annotation in the ImageNet dataset containing 3624 object categories as in [46].

This section also evaluates the proposed model on videos although it is designed primarily for individual images and does not exploit motion information during learning. Only a simple temporal smoothing post-processing step is introduced (see Section 3.5). YTO dataset [40] is a weakly annotated dataset composed of 10 object classes

3. Bayesian Joint Modelling for Object Localisation

in videos from YouTube. These 10 classes are a subset of the 20 VOC classes, which facilitate domain transfer experiments.

Features By default, the proposed method uses only a single appearance feature, namely SIFT to compare directly with most prior WSOL work which uses the same feature. Given an image j , the proposed model computes N_j 128-bin SIFT descriptors, regularly sampled every 5 pixels along both directions, and quantises them into a 2000-word codebook using K-means clustering. Differently to other bag-of-words (BoW) approaches [183, 184] which then discard spatial information entirely, the presented method then represents each image j by the list of N_j visual words and corresponding locations $\{x_i, l_{ai}, l_{bi}\}_{i=1}^{N_j}$ where $\{l_{ai}, l_{bi}\}$ are the coordinates of each word.

This thesis additionally extracts two more BoW features at the same regular grid locations to test the feature fusion performance. They are: (1) Colour-LAB: Colour provides complementary information to SIFT gradients. The proposed method quantises colour histograms into three channels (8,16,16) of LAB space and concatenate them to produce a 40 dimensional feature vector. Visual words are then obtained by quantising the feature space using K-means with $K=500$. (2) Local Binary Pattern (LBP) [222]: 52 bin LBP feature vectors are computed and quantised into a 500-bin histogram.

Settings and implementation details For the proposed model, the model set the foreground topic number K^{fg} to be equal to the number of classes, and $K^{bg} = 20$ for background topics. α is set to 0 or 1 as discussed in Section 3.4. and π^0 is initialised by Eq. 3.8 as described in Section 3.6. μ^0 is initialised with the central of the image area. Λ^0 is initialised from the half size of the image area. This thesis run Eqs. (3.4)-(3.6) for a fixed 100 VMP iterations. The localisation performance is measured using CorLoc [40, 46]: an object is considered to be correctly localised in a given image if the overlap between the localisation box and the ground-truth (any instance of the target class) is greater than 50%. The CorLoc accuracy is then computed as the percentage of correctly localised images for each target class. The same measure has been used in all methods compared in conducted experiments.

3.8.2 Comparison with State-of-the-art

3.8.2.1 Results on VOC Dataset

Competitors This thesis compares the proposed joint modelling approach to the following state-of-the-art competitors:

Deselaers et al. [42] A CRF-based multi-instance approach that localises object instances while learning object appearance. They report performance both with a single feature (GIST) and four appearance features (GIST, colour histogram, BoW of SURF, and HOG).

Pandey and Lazebnik [44] They adapt the FS deformable part-based models to address the WS localisation problem.

Siva and Xiang [90] A greedy search method based on Genetic Algorithm to localise the optimal object bounding box location against a costing function combining the object saliency, intra-class and inter-class cues.

Siva et al. NM [43] A simple Negative Mining (NM) approach which shows that inter-class is a stronger cue than the intra-class one when used properly.

Siva et al. OS [218] The NM approach above is extended to mine Objective Saliency (OS) information from a large corpus of unlabelled image. This can be considered as a hybrid of the object saliency approach in [69] and the NM work in [43].

Shi et al. [91] A ranking based transfer learning approach using an auxiliary dataset to score each candidate bounding box location in an image according to the degree of overlap with the unknown true location.

Zhu et al. [223] An unsupervised saliency guided approach to localise an object in a weakly labelled image in a multiple instance learning framework.

Tang et al. [46] An optimisation-centric approach that uses a convex relaxation of the MIL formulation.

Note that a number of the competitors [42, 43, 46, 90, 91] used an additional auxiliary dataset that the proposed model does not use. Objectness trained on auxiliary data was required by [42, 43, 46, 90, 91]. Although Shi *et al.* [91] evaluated all 20 classes, a randomly selected 10 were used as auxiliary data with bounding-boxes annotation. Pandey and Lazebnik [44] set aspect ratio manually and/or performed cropping on the obtained bounding-boxes. Note that Cabral *et al.* [193] also provides a mechanism for

3. Bayesian Joint Modelling for Object Localisation

recovering the location of objects in images by decoupling appearance descriptions of co-occurring classes. Unlike [193], which focus on the task of multi-label image classification, this work proposes an elaborately designed model for weakly supervised object localisation task.

Method	Initialisation			Refined by detector		
	6×2	14	20	6×2	14	20
Deselaers <i>et al.</i> [42]						
a. single feature	35	21	-	40	24	-
b. all four features	39	22	-	50	28	-
Pandey and Lazebnik [44] *						
a. before cropping	36.7	20.0	-	59.3	29.0	-
b. after cropping	43.7	23.0	-	61.1	30.3	-
Siva and Xiang [90]	40	-	28.9	49	-	30.4
Siva <i>et al.</i> NM [43]	37.1	-	29.0	46	-	-
Siva <i>et al.</i> OS [218]	42.4	-	31.1	55	-	32.0
Shi <i>et al.</i> [91] +	39.7	-	32.1	-	-	-
Zhu <i>et al.</i> [223]	-	-	-	-	31	-
Tang <i>et al.</i> [46]	39	-	-	-	-	-
Cinbis <i>et al.</i> [224]		-	-	-	-	38.8
Our-Sampling	50.8	32.2	34.1	65.5	33.8	36.2
Our-Gaussian	51.5	30.5	31.2	66.1	32.5	33.4
Our-Sampling+prior	51.2	33.4	36.1	65.9	35.4	38.3
Our-Gaussian+prior	51.8	31.1	33.5	66.7	33.0	35.8

Table 3.1: Comparison with state-of-the-art competitors on the three variations of the PASCAL VOC 2007 dataset. * Requires aspect ratio to be set manually. + Require 10 out of the 20 classes fully annotated with bounding-boxes and used as auxiliary data.

Initial localisation Table 3.1 shows that for the initial annotation accuracy the proposed model consistently outperforms all competitors over all three VOC variants, sometimes by big margins. This is mainly due to the unique joint modelling approach taken by the proposed method, and its ability to integrate prior spatial and appearance knowledge in a principled way. Note that the prior knowledge is either based on first principle (spatial and appearance) or computed from the data without any additional human intervention (appearance). The proposed two object localisation methods (Our-Sampling and Our-Gaussian) vary in performance over different-sized datasets. Our-Gaussian performs better in the relatively simple datasets (6×2) where most images

3. Bayesian Joint Modelling for Object Localisation

contain only one object, because the proposed Gaussian location model can compact objects easily in this case. In contrast, Our-Sampling is better in the more complicated situation (20 classes) where many objects co-existing in one image is more common.

Refined by detector After the initial annotation of the weakly labelled images, a conventional strong object detector can be trained using these annotations as ground truth. The trained detector can then be used to iteratively refine the object location. This thesis follows [44, 90] in exploiting a deformable part-based model (DPM) detector¹ [16] for one iteration to refine the initial annotation. Table 3.1 shows that again the proposed model outperforms almost all competitors by a clear margin for all three datasets (see the appendix A for more detailed per-class comparisons). Very recently, [224] achieved similar performance by training a multi-instance SVM with a more powerful fisher vector based representation.

With appearance similarity prior As described before, the proposed framework can exploit the appearance similarity prior across classes. Although the actual appearance similarity between classes is hard to calculate, the proposed model can approximate it by computing the relatedness using WordNet semantic tree [225]. Figure 3.4 shows the pairwise relatedness among 20 classes, which is generated using the Lin distance of [200]. The diagonal of the matrix verifies that classes are most similar to themselves. Leaf nodes (blue) correspond to the classes of VOC-20. Classes that inherit from the same subtree should show more similar appearance. A pairwise similarity matrix is then calculated from the tree structure and used to correlate their appearance as explained in Section 3.6. The bottom two rows of Table 3.1 show the localisation accuracy with the appearance similarity prior. It clearly shows that the prior improves the performance of both variants of the proposed model for all experiments. It is interesting to note that the performance is improved more on VOC-20 than VOC-6 \times 2. This is because there is more opportunity to share related appearance as the number of classes increases. Categories in 6 \times 2 are generally more dissimilar, so there is less benefit to the correlation.

What has been learned Figure 3.5 gives examples of the localisation results and illustrates what has been learned for the foreground object classes. For the latter, this section shows the response of each learned object topic (i.e. the posterior probability

¹Version 3.0 is used for fair comparison against most published results obtained using the same version.

3. Bayesian Joint Modelling for Object Localisation

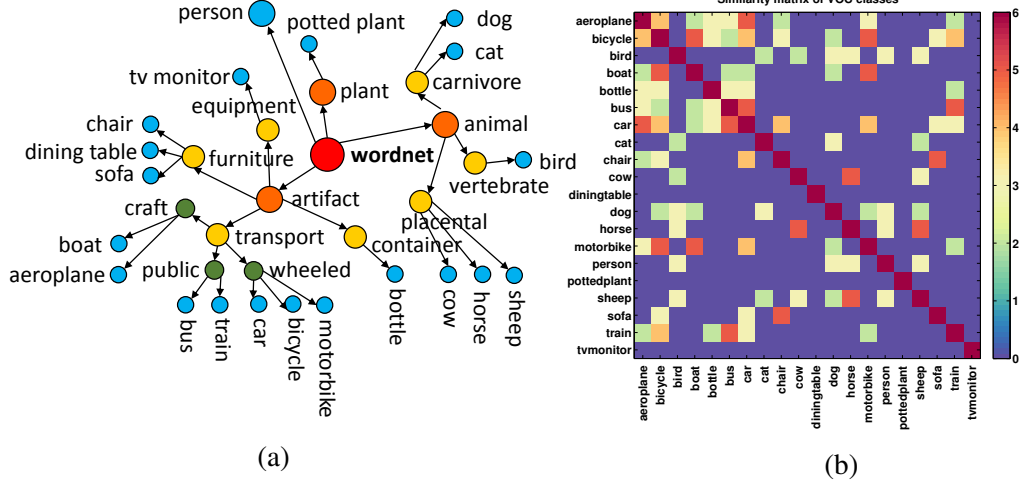


Figure 3.4: (a) A hierarchical structure of the 20 PASCAL VOC classes using WordNet. (b) The class similarity matrix.

of the topic given the visual word) as a gray-level image, or heat map (the brighter, the higher probability that the object is present at each image location). These examples show that the foreground topics indeed capture what each object class looks like and can distinguish it from the background and between different object classes. For instance, Figure 3.5(c) shows that the motorbike heat map is quite accurately selective, with minimal response obtained on the other vehicular clutter. Figure 3.5(e) indicates how the Gaussian can sometimes give a better bounding box. The opposite is observed in Figure 3.5(f) where the single Gaussian assumption is not ideal when the foreground topic has less a compact response. Selectivity is illustrated by Figure 3.5(c,d), Figure 3.5(h,i) and Figure 3.5(g,k), which show the same images, but with detection results for different co-occurring objects. In each case, the relevant object has been successfully selected while “explaining away” the potentially distracting alternative. The presented method may fail if the background clutter or objects of no interest dominates the image (Figure 4(l,m,u)). For example, in Figure 3.5(l), a bridge structure resembles the boat in Figure 3.5(a) resulting strong response from the boat topic, whilst the actual boat, although picked up, is small and overwhelmed by the false response.

A key strength of the proposed framework is explicit modelling of background without any supervision. This allows background pixels to be explained, reducing

3. Bayesian Joint Modelling for Object Localisation

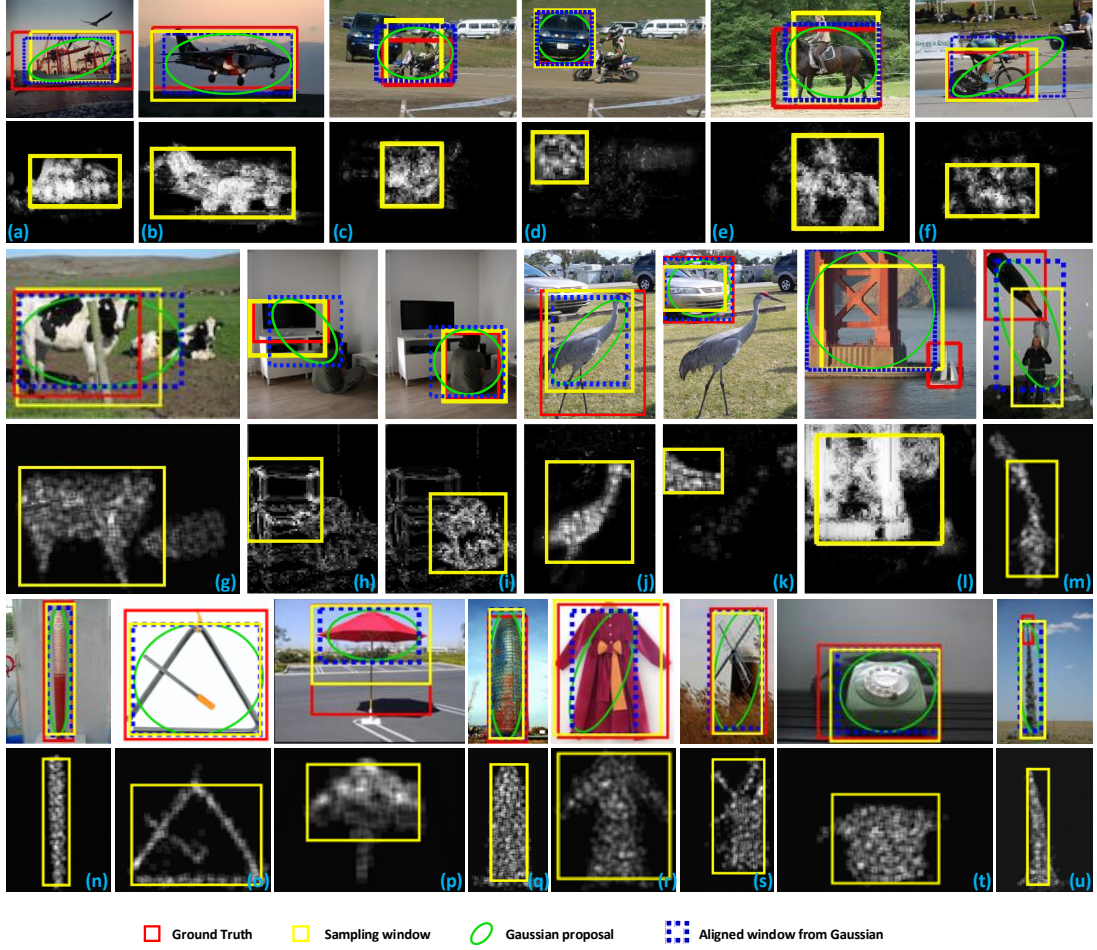


Figure 3.5: Top row in each subfigure: examples of object localisation using our-sampling and our-Gaussian. Bottom row: illustration of what is learned by the object (foreground) topics via heat map (brighter means object is more likely). The first four rows show some examples of PASCAL VOC and last two rows are selected from ImageNet.

confusion with foreground objects and hence improving localisation accuracy. This is illustrated in Figure 3.6 via plots of the background topic response (heat map). It illustrates qualitatively that some background topics are often correlated with common semantic background components such as sky, grass, road and water, despite none of these being annotated.

Weakly supervised detector The ultimate goal of WSOL is to learn a WS detector. This is achieved by feeding the localised objects into an off-the-shelf detector training

3. Bayesian Joint Modelling for Object Localisation

model. The deformable part based model (DPM) in [16] is used and this WS detector is compared against a FS one with the same DPM model (version 3.0). Specifically, Table 3.2 compares the mean average precision (mAP) of detection performance on both VOC-6×2 and VOC-20 test datasets among previous reported WS detector results, ours and the FS detector [16]. Due to the better localisation performance on the WS training images, the proposed approach is able to reduce the gap between the WS detector and the FS detector. The detailed per-class result is included in the appendix A and it shows that for classes with high localisation accuracy (e.g. bicycle, car, motorbike, train), the WS detector is often as good as the FS one, whilst for those with very low localisation accuracy (e.g. bottle and pottedplant), the WS detector fails completely.

3. Bayesian Joint Modelling for Object Localisation

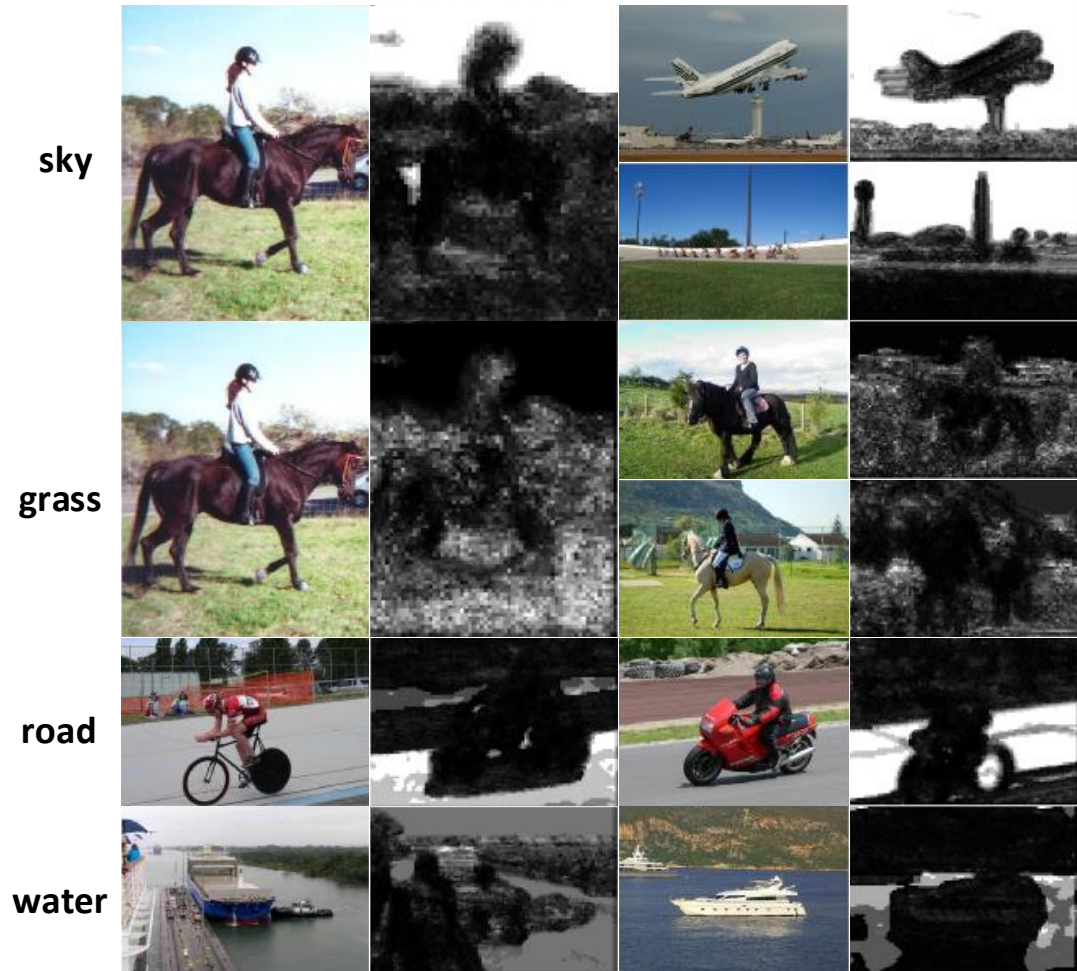


Figure 3.6: Illustration of the learned background topics.

Data	Deselaers [42]	Pandey [44]	Siva [90]	Ours	Fully Supervised [16]
6×2	21	20.8	-	26.1	33.0
20	-	-	13.9	17.2	26.3

Table 3.2: Performance of strong detectors trained using annotations obtained by different WSOL methods

3. Bayesian Joint Modelling for Object Localisation

3.8.2.2 Results on ImageNet Dataset

Method	Initialisation
Alexe <i>et al.</i> [69]	37.4
Tang <i>et al.</i> [46]	53.2
Our-Sampling	57.6

Table 3.3: Initial annotation accuracy on ImageNet dataset

Table 3.3 shows the initial annotation accuracy of different methods for the much larger 3624-class ImageNet dataset. Note that the result of Alexe *et al.* [69] is taken from the Table 4 in [46]. Although the annotation accuracy could be further improved by training an object detector to refine the annotation as shown in Table 2, this step is omitted in current experiment as none of the competitors attempted it. For such a large scale learning problem, loading all the image features into the memory is a challenge for the proposed joint learning method. A standard solution is taken, that is, to process in batches of 100 classes. Joint learning is performed within each batch but not across batches; the proposed model is thus not used to its full potential. Table 3.3 shows that the proposed method achieves the best result (57.6%). Note that [69] is a very simple baseline as it simply takes the top-scoring objectness box. Recently more sophisticated transfer-based techniques [38] and [226] were evaluated on ImageNet. But their results were obtained on a different subset of ImageNet, thus not directly comparable here.

To investigate the effect of the similarity prior in this larger dataset, the presented method randomly chooses 500 small (containing around 100 images each) leaf-node classes from ImageNet for joint-learning with an inter-class similarity prior. This was the largest dataset size that could simultaneously fit in the memory of the used platform¹. Performing joint learning with inter-class correlation on this ImageNet subset, the proposed model achieves 58.8% annotation accuracy on the 500 classes compared to 55.4% without using the similarity prior.

¹The proposed learning algorithm could potentially be modified to process all 3624 classes in batches.

3.8.2.3 Results on YouTube-object Dataset

The main competitors on YTO are [40] and [93]. Prest *et al.* [40] first performed spatio-temporal segmentation of video into a set of 3D tubes, and subsequently searched for the best object location. Very recently, [93] simultaneously localised objects of the same class across a set of video clips (co-localisation) with the Frank-Wolfe Algorithm. Note that there are some recently published studies on WS object segmentation from video [213]. This is not directly comparable as they did not report results based on the standard YTO bounding-box annotations. Two variants of the proposed model are compared here: Our-sampling is the method evaluated above for individual images. Used here, it ignores the temporal continuity of the video frames in a video. Our-smooth is the simple extension of Our-sampling for video object localisation. As described in Section 3.5, temporal information is used to enforce a smooth change of object location over consecutive frames. The way temporal information is exploited is thus much less elaborate than that in [40]. For all methods compared, this section evaluates localisation performance on the key frames which are provided with ground truth labels by [40].

Table 3.4 shows that even without using any temporal information and operating on key frames only, Our-sampling outperforms the method in [40]. Our-Smooth further improves the performance and the localisation accuracy of 32.2% is very close to the upper bound result (34.8%) suggested by [40], which is the best possible result from oracle tube extraction. Figure 3.7 shows some examples of video object localisation using Our-Smooth. It is worth to note that all these results have been exceeded (50.1% accuracy) recently by a model purposefully designed for video segmentation [227], which performed much more intensive spatio-temporal modelling and used superpixel segmentation within each frame and motion segmentation across frames.

3. Bayesian Joint Modelling for Object Localisation

Categories	[40]	[93]	Our-Sampling	Our-Smooth	[227]
aeroplane	51.7	27.5	40.6	45.9	65.4
bird	17.5	33.3	39.8	40.6	67.3
boat	34.4	27.8	33.3	36.4	38.9
car	34.7	34.1	34.1	33.9	65.2
cat	22.3	42.0	35.3	35.3	46.3
cow	17.9	28.4	18.9	22.1	40.2
dog	13.5	35.7	27.0	27.2	65.3
horse	26.7	35.6	21.9	25.2	48.4
motorbike	41.2	22.0	17.6	20.0	39.0
train	25.0	25.0	32.6	35.8	25.0
Average	28.5	31.1	30.1	32.2	50.1

Table 3.4: Performance comparison on YouTube-object

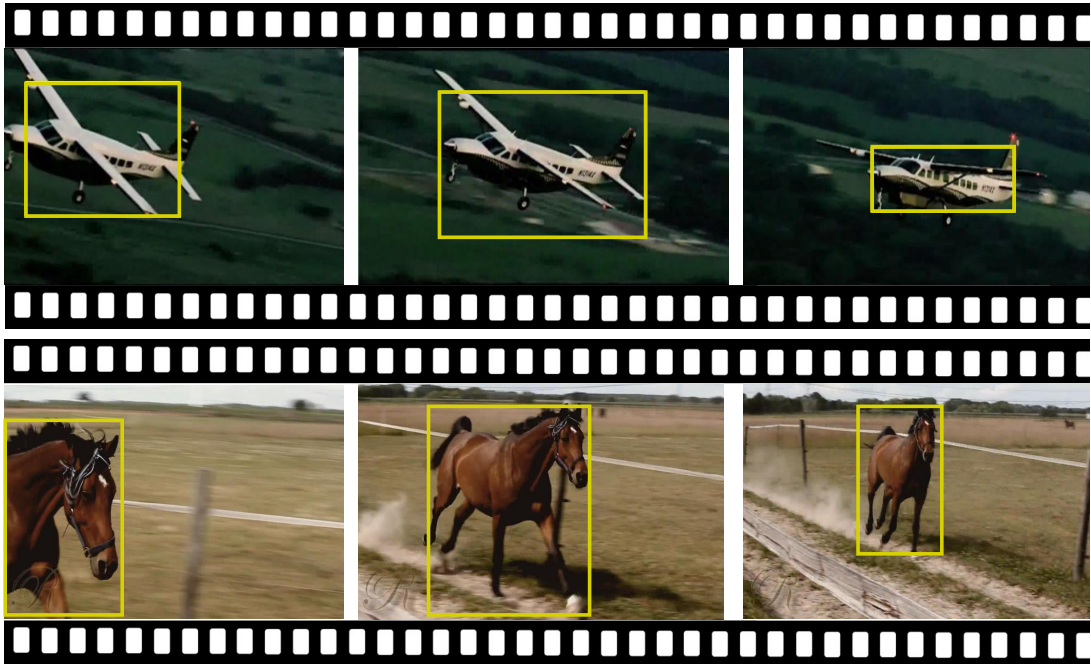


Figure 3.7: Examples of video object localisation

3.8.3 Bayesian Domain Adaptation

This thesis next evaluates the potential of the proposed model for WS cross-domain transfer learning using the YTO and VOC07-10 as the two domains (the designed method chooses the same 10 classes from the VOC07-20 as in YTO). One domain contains continuous and highly varying video data, and the other contains high resolution but cluttered still images. This thesis considers following two non-transfer baselines:

YTO, VOC The first baseline is the original performance on YTO and VOC07-10 classes, solely using target domain data. *YTO* is exactly the same as Our-Sampling described in Section 3.8.2.3, while *VOC* is trained with 10 classes from VOC07-20 using the same setting described in Section 3.8.2.1.

All→YTO, All→VOC The second baseline simply combines the training data of YouTubeObject and VOC. One model trained with these two domains' data is used to localise object on YouTubeObject ($A \rightarrow Y$) and VOC07-10 ($A \rightarrow V$).

The proposed model considers two directions of knowledge transfer between YTO and VOC07-10, and compare the above baselines with the presented domain adaptation method: $V \rightarrow Y$ is initialised with an appearance prior transferred from VOC07-10, and adapted on the YTO data. On the contrary, $Y \rightarrow V$ adapts the YTO appearance prior to VOC07-10. Table 3.5 shows that the proposed Bayesian domain adaption method performs better than the baselines on both YTO and VOC07-10. In contrast, the standard combination ($A \rightarrow Y$ and $A \rightarrow V$) shows little advantage over solely using target domain data. Note that unlike prior studies of video→image [40] or image→video [97] that adapt detectors with fully labelled data, the interested task is to adapt weakly labelled data.

This thesis also varies the amount of target domain data and evaluate its effect on the domain transfer performance. Figure 3.8 shows that the presented model provides a bigger margin of benefit given less target domain data. This can be easily understood because with a small quantity of training examples there is insufficient data to learn the object appearance well and the impact of the knowledge transfer is thus more significant.

3. Bayesian Joint Modelling for Object Localisation

Categories	YTO			VOC		
	Y	A→Y	V→Y	V	A→V	Y→V
aeroplane	40.6	40.8	45.8	57.5	58.1	58.7
bird	39.8	40.3	38.8	29.8	30.5	33.7
boat	33.3	33.4	38.8	28.0	27.9	29.0
car	34.1	33.9	33.6	39.1	39.1	44.4
cat	35.3	35.3	38.8	59.0	59.3	58.6
cow	18.9	19.0	27.7	36.7	36.9	38.9
dog	27.0	27.1	26.7	46.5	47.4	48.3
horse	21.9	22.1	26.1	53.2	53.5	55.5
motorbike	17.6	17.9	17.5	55.6	55.2	58.1
train	32.6	32.6	36.2	54.7	54.5	56.3
Average	30.1	30.2	33.0	46.0	46.2	48.1

Table 3.5: Cross-domain transfer learning results. The proposed transfer learning strategy ($V \rightarrow Y$ and $Y \rightarrow V$) shows substantial improvements over the standard combinations ($A \rightarrow Y$ and $A \rightarrow V$) and the baselines (Y and V).

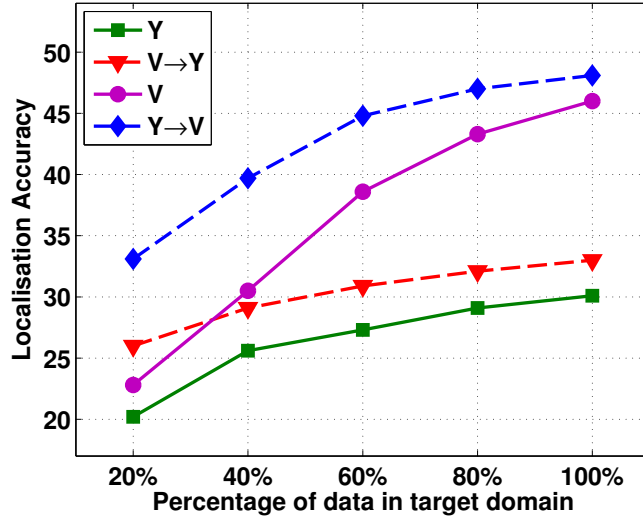


Figure 3.8: Domain adaptation provides more benefit with fewer target domain samples. Specifically, the presented model ($V \rightarrow Y$ and $Y \rightarrow V$) provides a bigger margin of benefit given less target domain data.

3.8.4 Semi-supervised Learning

One important advantage of the proposed model is the ability to utilise unlabelled data to further reduce the manual annotation requirements. To demonstrate this the proposed model randomly selects 10% of the *VOC07-6×2* data as weakly labelled training data, and then varies the additional unlabelled data used. Note that 10% labelled data corresponds to around only 5 *weakly labelled images per class* for the *VOC07-6×2* dataset, which is significantly less than what any previous method has exploited. Two evaluation procedures are considered: (i) Evaluating localisation performance on the initially annotated 10% (standard WSOL task); and (ii) WSOL performance on the held out *VOC07-6×2* test set¹. The latter corresponds to an online application scenario where the localisation model is trained on one database and needs to be applied online to localise objects in incoming weakly labelled images. This thesis varies the additional data across a combination of four conditions: (1) *6R*: add the remaining 90% of data for the 6 target classes but without labels, (2) *100U*: add all images from 100 unrelated ImageNet classes without labels, (3) *6R + 100U*: add both of the above. There are two questions to answer: Whether the model can exploit the related data when it comes without labels (*6R*), and whether it can avoid being confused by a vast quantity of unrelated data (*100U*).

The results are shown in Table 3.8, where the ratio of relevant to irrelevant data in the additional unlabelled samples is shown in the second column. From the results, this thesis can draw the following conclusions: (1) As expected, the model performs poorly with little data (10%L). However it improves significantly with some relevant but unlabelled data (the standard SSL setting, 10%L+6R). Moreover, this SSL result is almost as good as when all the data are labelled (100%L). (2) If *only* irrelevant data are added to the small labelled seed, not only does the performance not degrade, but it increases noticeably (10%L vs. 10%L+100U). (3) If both relevant and irrelevant data are added – corresponding to the realistic scenario where an automatic process gathers a pool of potentially relevant data which, without any screening, will be a mix of relevant and irrelevant data to the target problem. In this case the performance improves to not far off the fully annotated case (10%L vs. 10%L+6R+100U vs. 100%L). As expected,

¹To localise objects in a test image, the proposed model only needs to iterate Eqs. (3.4)-(3.5) instead of (3.4)-(3.6). That is, the object appearance is considered fixed and does not need to be updated. This both reduces the cost of each iteration and also makes convergence more rapid.

3. Bayesian Joint Modelling for Object Localisation

the performance of 10%L+6R+100U is weaker than 10%L+6R – if one manually goes through the unlabelled data and removes the irrelevant ones and leave only the relevant ones, it would certainly benefit the model. But it is noted that the decrease in performance is small (47.1% to 43.5%). (4) If the irrelevant data are added to the fully annotated dataset, the performance improves slightly (100%L vs. 100%L+100U), which shows that the proposed model is robust to this potential distraction from the large amount of unlabelled and irrelevant data. This is expected in SSL, which typically benefits only when the amount of labelled data are small. These results show that the proposed approach has good promise for effective use in realistic scenarios of learning from only few weak annotations and a large volume of only partially relevant unlabelled data. This is illustrated visually in Figure 3.9, where unlabelled data helps to learn a better object model. Finally, the similarly good results on the held-out test set verify that the proposed model is indeed learning a good generalisable localisation mechanism and is not over-fitted to the training data.

VOC07-6 \times 2		Data for Localisation	
Data for Training	ratio of R:U	10%L	Test set
10%L	-	27.1	28.0
10%L+6R	1	47.1	42.3
10%L+100U	0	35.8	32.4
10%L+6R+100U	0.04	43.5	38.1
100%L	-	50.3	46.2
100%L+100U	0	50.7	47.5

Table 3.6: Localisation performance of semi-supervised learning. Unlabelled data helps to learn a better object model.

3.8.5 Computational Cost

The proposed model is efficient both in learning and inference, with a complexity $\mathcal{O}(NMK)$ for N images, M observations (visual words) per image, and K classes. The experiments were done on a 2.6GHz PC with a single-threaded Matlab implementation. Training the model on all 5,011 VOC07 images required 3 hours and a peak

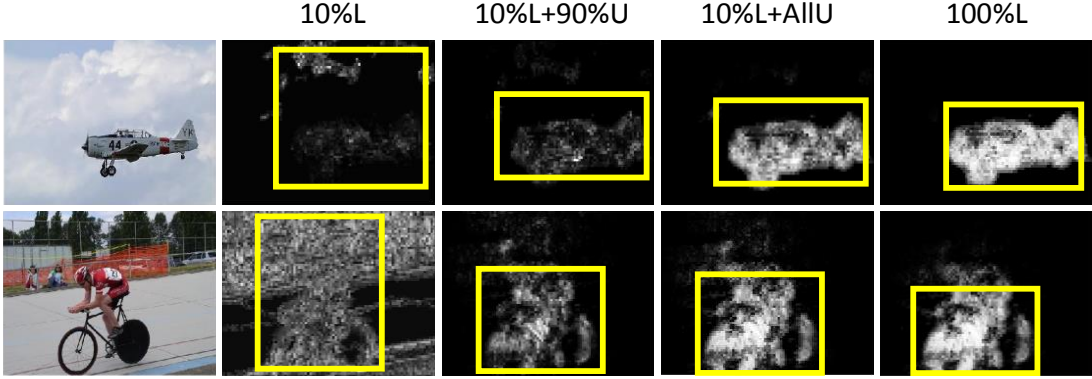


Figure 3.9: Unlabelled data improves foreground heat maps.

of 6 GB of memory to learn a joint model for 20 classes. The proposed Bayesian topic inference process not only enables prior knowledge to be used, but also achieves 10-fold improvements in convergence time compared to EM inference used by most conventional topic models with point-estimated Dirichlet topics. Online inference of a new test image took about 0.5 seconds. After model learning, for object localisation in training images, direct Gaussian localisation is effectively free and heat-map sampling took around 0.6 seconds per image. These statistics compare favourably to alternatives: [42] reported 2 hours to train 100 images; while the Matlab implementations of [43], [90] and [186] took 10, 15 and 20 hours respectively to localise objects for all 5,011 images.

3.9 Summary

This chapter has presented an effective and efficient model for WSOL. The proposed approach surpasses the performance of all prior methods and obtains state-of-the-art results thanks to three key properties: 1) jointly learning a model for all classes, 2) a Bayesian formulation, and 3) an explicit notion of the spatial location of an object. Moreover, it is also possible to perform SSL to obtain an effective localiser with only a fraction of the annotated training data. Moreover, the unlabelled data need not even be sanitised for relevance to the target classes. In this study the proposed model only used simple top-down cues via the proposed Bayesian priors; however this formulation has

3. Bayesian Joint Modelling for Object Localisation

great potential to enable more scalable learning through cross-class and cross-domain transfer via priors [34, 38, 91]. These contributions bring us significantly closer to the goal of scalable learning of strong models from weakly-annotated non-purpose collected data on the Internet.

However, detecting objects cannot tell us the whole story in the image content. One real world image is more than “what objects are in the image”, but also “how do they look like”. Object detection itself is not sufficient to describe image contents. Exploring attribute is hence desired to provide rich descriptive explanations for visual content. In addition, this framework is not ready for modelling complex visual scenes relation from the weak label. For example, an unlimited number of attributes can co-exist on one object. These are not considered in this chapter. The next chapter addresses these limitations by formulating a non-parametric Bayesian model. The proposed approach aims to model and learn the appearance of object and attribute classes as well as their association, where only given weak image-level annotations of objects and attributes without locations or associations between them. The new method also allows an infinite number of factors to model the complex relation among objects, attributes and background clutters.

Chapter 4

Weakly Supervised Learning of Objects, Attributes and their Associations

4.1 Overview

As discussed in Section 1.1, a real world image contains complex information such as objects, attributes and object-attribute associations. The preceding chapter has described a Bayesian joint topic model to jointly learn all object classes and image backgrounds for object detection. This approach faces three main limitations: First, object detection only provides partial information of image contents. To describe the object appearance more semantically, visual attribute is a powerful cue for representing generic objects. The previous chapter mainly focused on detecting objects in given images. Second, the approach described in chapter 3 addresses WS object detection problem by solving two sub-problems simultaneously: 1) locating the objects of interest in each training image, 2) training an object detector based on the annotation results from 1). It indicates that a separate classic object detection model is required to perform the whole process. Therefore, a holistic approach is preferred to deal with a new test image directly. Third, it is non-trivial to model complicated visual relation by chapter 3's approach directly, especially when considering objects and attributes simultaneously. This is because the parametric model is hard to explain a large number of factors (e.g.

4. Weakly Supervised Learning of Objects, Attributes and their Associations

objects, attributes, and infinite latent factors) from image-level annotations. The fact that a unlimited number of attributes can describe the same instance (e.g. one object or background region) is contrary to the assumption of prior distribution in the previous chapter. These limitations prevent the previous model from generalising to discover complex semantics and understand image content comprehensively.

In order to overcome these limitations, a new approach is proposed in this chapter to model object, attribute, and their associations from weakly labelled images. In particular, when humans describe images they tend to use combinations of nouns and adjectives, corresponding to objects and their associated attributes respectively. To generate such a description automatically, one needs to model objects, attributes and their associations. Conventional methods require strong annotations of object/attribute locations, making them less scalable. The proposed model aims to exploit weakly labelled images, such as those widely available on media sharing sites (e.g. Flickr), where only image-level labels (either object or attributes) are given, without their locations and associations. This is achieved by introducing a novel WS non-parametric Bayesian model. Once learned, given a new image, the proposed model can describe the image, including objects, attributes and their associations, as well as their locations and segmentation (see Figure 4.1). Extensive experiments on benchmark datasets demonstrate that the proposed WS model performs at par with strongly supervised models on tasks such as image description and retrieval based on object-attribute associations.

Modelling weakly labelled images using the proposed framework provides a number of benefits: (i) By jointly learning multiple objects, attributes and background clutter in a single framework, ambiguity in each is explained away by knowledge of the other. (ii) The infinite number of factors provided by the non-parametric Bayesian framework allows structured background clutter of unbounded complexity to be explained away. (iii) A sparse binary latent representation of each patch allows an unlimited number of attributes to co-exist on one object. The aims and capabilities of the proposed approach are illustrated schematically in Figure 4.1, where weak annotation in the form of a mixture of objects and attributes is transformed into objects and attributes associations with locations.

The remainder of this chapter is organised as follows: Section 4.2 presents the proposed WS-SIBP. This includes the explanation of model learning and inference for test data. The various applications of the learned model are also discussed in this

4. Weakly Supervised Learning of Objects, Attributes and their Associations

section. Experimental results are reported in Section 4.3. Finally, a summary is given in Section 4.4.

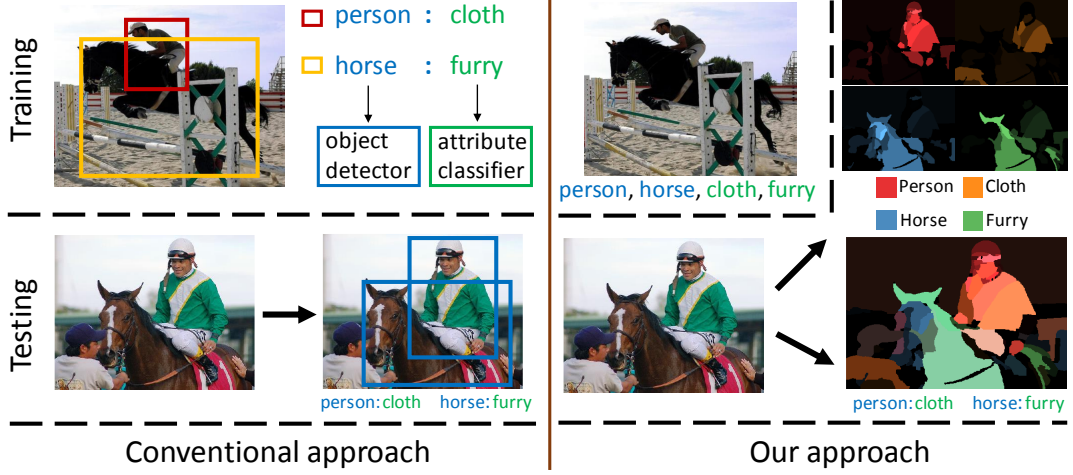


Figure 4.1: Comparing the proposed WS approach to object-attribute association learning to the conventional strongly supervised approach.

4.2 Weakly Supervised Stacked Indian Buffet Process

4.2.1 Image Representation

Given a set of images labelled with image-level object and attribute labels without explicitly specifying which attribute is associated with which object, the presented method aims to learn a model that, given a unseen new image, segments each object in an image and assign both object and attribute labels to it. As in most previous semantic segmentation works, the proposed approach first decomposes each image into superpixels which are over-segmented image patches that typically contain object parts. The problem of joint object and attribute annotation and segmentation thus boils down to multi-label classification of each superpixel, from which various tasks such as image-level annotation, object-attribute association, and object segmentation can be performed.

Each image j in a training set is decomposed into N_j super-pixels using a recent hierarchical image segmentation algorithm [105]¹. Each segmented superpixel is rep-

¹The proposed model set the segmentation threshold in the algorithm to 0.1 to obtain a single over-segmentation from the hierarchical segmentations for each image

4. Weakly Supervised Learning of Objects, Attributes and their Associations

resented using two types of normalised histogram features: SIFT and Color. (1) SIFT: the feature is extracted by regular grid (every 5 pixels) colorSIFT [228] at four scales. A 256 component GMM model is constructed on the collection of ColourSIFTs from all images. The presented method computes Fisher Vector + PCA for all regular points in each superpixel following [229]. The resulting reduced descriptor is 512-D for every segmented region. (2) Colour: the image is converted to quantised LAB space $8 \times 8 \times 8$. A 512-D color histogram is then computed for each patch. The final normalised 1024-D feature vector concatenates SIFT and Colour features together.

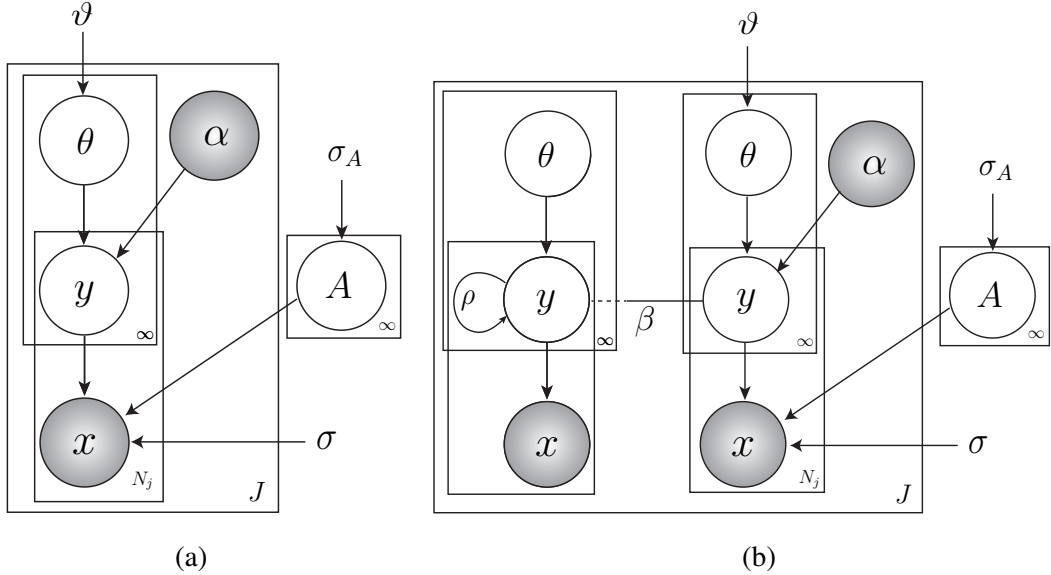


Figure 4.2: The probabilistic graphical model representing the proposed WS-MRF-SIBP. Shaded nodes are observed.

4.2.2 Model Formulation

This chapter proposes a non-parametric Bayesian model that learns to describe images composed of super-pixels from weak image-level object and attribute annotation. In the proposed model, each superpixel is associated with an infinite latent factor vector indicating if it corresponds to (an unlimited variety of) unannotated background clutter, or an object of interest, and what set of attributes are possessed by the object. Given a set of images with weak labels and segmented into super-pixels, the proposed model needs to learn: (i) which are the unique superpixels shared by all images with a

4. Weakly Supervised Learning of Objects, Attributes and their Associations

particular label, (ii) which superpixels correspond to unannotated background, and (iii) what is the appearance of each object, attribute and background type. Moreover, since multiple labels (attribute and object) can apply to a single superpixel, the model needs to disambiguate which aspects of the appearance of the superpixel are due to each of the (unknown) associated object and attribute labels. To address the WSL tasks the presented method builds on the IBP [189] and first introduce in Section 4.2.2.1 a WS-SIBP to model data represented as bags (images) of instances (superpixels) with bag-level labels (image annotations). This is analogous to the notion of documents in topic models [179]. Furthermore, to fully exploit spatial and inter-factor correlation, two types of MRFs are integrated (see Section 4.2.2.2), resulting in the full model termed WS-MRF-SIBP as illustrated in Figure 4.2.

4.2.2.1 WS-SIBP

This model aims to associate each image/superpixel with a latent factor vector whose elements will correspond to objects, attributes and/or unannotated attribute/background present in that image/superpixel. Let each image j represented as bags of superpixels $\mathbf{X}^{(j)} = \{\mathbf{X}_{i\cdot}^{(j)}\}$, where the notation $\mathbf{X}_{i\cdot}$ means the vector of row i in matrix \mathbf{X} , i.e. the 1024-D feature vector representing the i -th superpixel, and $i \in 1 \dots N_j$. Assuming there are K_o object categories and K_a attributes in the provided image-level annotations, they are represented by the first $K_{oa} = K_o + K_a$ latent factors. In addition, an unbounded number of further factors are available to explain away background clutter in the data, as well as discover undefined latent attributes. At training time, the model assumes a binary label vector $\boldsymbol{\alpha}^{(j)}$ for objects and attributes is provided for each image i . So $\alpha_k^{(j)} = 1$ if attribute/object k is present, and zero otherwise. Also $\alpha_k^{(j)} = 1$ for all $k > K_{oa}$. That is, without any labels, it assumes all background/latent attribute types can be present. With these assumptions, the generative process (illustrated in Figure 4.2) for the image i is as follows:

For each latent factor $k \in 1 \dots \infty$:

1. Draw an appearance distribution mean $\mathbf{A}_k \sim \mathcal{N}(0, \sigma_A^2 \mathbf{I})$.

For each image $j \in 1 \dots J$:

4. Weakly Supervised Learning of Objects, Attributes and their Associations

1. Draw a sequence of i.i.d. random variables $b_1^{(j)}, b_2^{(j)} \dots \sim \text{Beta}(\vartheta, 1)$,
2. Construct an image prior $\theta_k^{(j)} = \prod_{t=1}^k b_t^{(j)}$,
3. Input weak annotation $\alpha_k^{(i)} \in \{0, 1\}$,
4. For each superpixel $i \in 1 \dots N_j$:
 - (a) Sample state of each latent factor k : $y_{ik}^{(j)} \sim \text{Bern}(\theta_k^{(i)} \alpha_k^{(i)})$,
 - (b) Sample superpixel appearance: $\mathbf{X}_{j\cdot}^{(i)} \sim \mathcal{N}(\mathbf{Y}_{j\cdot}^{(i)} \mathbf{A}, \sigma^2 \mathbf{I})$.

where \mathcal{N} , Bern and Beta respectively correspond to Normal, Bernoulli and Beta distributions with the specified parameters. The Beta-Bernoulli and Normal-Normal conjugacy are chosen because they allow more efficient inference. ϑ is the prior expected sparsity of annotations and σ^2 is the prior variance in appearance for each factor.

This generative process encodes the assumptions that the available factors for each superpixel are determined by the image level labels if given (generative model for \mathbf{Y}); and that multiple factors come together to explain each superpixel (generative model for \mathbf{X} given \mathbf{Y}).

Joint probability: Denote hidden variables by $\mathbf{H} = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(J)}, \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(J)}, \mathbf{A}\}$, J images in a training set by $\mathbf{X} = \{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(J)}\}$, and parameters by $\boldsymbol{\Pi} = \{\vartheta, \sigma_A, \sigma, \boldsymbol{\alpha}\}$. Then the joint probability of the variables and data given the parameters is:

$$\begin{aligned}
 p(\mathbf{H}, \mathbf{X} | \boldsymbol{\Pi}) = & \prod_{i=1}^J \left(\prod_{k=1}^{\infty} \left(p(\pi_k^{(j)} | \vartheta) \prod_{i=1}^{N_j} p(y_{ik}^{(j)} | \pi_k^{(j)}, \alpha_k^{(j)}) \right) \right. \\
 & \cdot \prod_{i=1}^{N_j} p(\mathbf{X}_{i\cdot}^{(j)} | \mathbf{Z}_{i\cdot}^{(j)}, \mathbf{A}, \sigma) \Big) \\
 & \cdot \prod_{k=1}^{\infty} p(\mathbf{A}_k | \sigma_A^2).
 \end{aligned} \tag{4.1}$$

4. Weakly Supervised Learning of Objects, Attributes and their Associations

Learning in the presented model (detailed in Section 4.2.3) aims to compute the posterior that $p(\mathbf{H}|\mathbf{X}, \mathbf{\Pi})$ for: disambiguating and localising all the annotated ($\alpha^{(j)}$) objects and attributes among the superpixels (inferring $\mathbf{Y}_i^{(j)}$), inferring the attribute and background prior for each image (inferring $\theta^{(j)}$), and learning the appearance of each factor (inferring \mathbf{A}_k).

4.2.2.2 WS-MRF-SIBP

Now the presented method generalises the i.i.d. model WS-SIBP to WS-MRF-SIBP by introducing two types of factor correlation.

Spatial MRF for smoothing factor assignments across superpixels: As mentioned above, each superpixel’s latent factors are so far drawn from the image prior θ_k^j – independently of their neighbours (Eq. (4.1)). Thus spatial structure is ignored in WS-SIBP, although adjacent superpixels are known to be strongly correlated in real images [190]. Inspired by the successful use of random fields for capturing the spatial coherence of image region labels [190, 191, 192], the presented approach introduces a MRF with connections between spatially adjacent nodes (superpixels). Specifically, the following MRF potential Φ [190, 220] is introduced to the generative process for \mathbf{Y} to correlate the superpixel factors drawn in image j spatially:

$$\Phi(\mathbf{Y}_{\cdot k}^{(j)}) = \exp \sum_{i, m \in N_j} \beta \mathbf{I}(y_{ik}^{(j)} = y_{mk}^{(j)}), \quad (4.2)$$

where $i, m \in N_j$ enumerates node pairs i, m that are spatial neighbours in image j . The indicator function \mathbf{I} returns 1 when its argument is true, i.e., when neighbouring superpixels have the same assignment for factor k . β is the coupling strength parameter of the MRF, which controls how likely they have the same label *a priori*. The initial WS-SIBP formulation can be obtained by setting $\beta = 0$. Spatial MRF is encoded for all given K_{oa} and newly discovered factors.

Factorial MRF within superpixel: Although individual factors are now correlated spatially, the current method does not yet model any inter-factor co-occurrence statistics within a single superpixel (as in most other MRF applications [190, 191]). However, exploiting this information (e.g., *person* superpixels more likely to share attribute *clothing*, than *metallic*) is important, especially in the ambiguous WSL setting. To

4. Weakly Supervised Learning of Objects, Attributes and their Associations

represent these inter-factor correlations, the proposed approach introduces a factorial MRF Θ via the following potential on the generative process for \mathbf{Y} :

$$\Theta(\mathbf{Y}_{i.}^{(j)}) = \exp \prod_{k,l}^{\infty} \Theta(y_{ik}^{(j)}, y_{il}^{(j)}) \quad (4.3)$$

$$\Theta(y_{ik}^{(j)}, y_{il}^{(j)}) = \begin{cases} 0 & \text{if } k = l \\ \rho \mathcal{M}(k, l) & \text{otherwise,} \end{cases} \quad (4.4)$$

where ρ controls the importance of the factorial MRF, and matrix \mathcal{M} encodes inter-factor correlations. In the traditional strongly-supervised scenario, \mathcal{M} can be trivially learnt from the fully labelled annotations. In the WSL scenario, \mathcal{M} cannot be determined directly. It will be discussed about how to learn \mathcal{M} in Section 4.2.3.

WS-MRF-IBP Prior: Overall, combining the two MRFs, the latent factor prior

$$p(\mathbf{Y}^{(j)} | \boldsymbol{\theta}^{(j)}, \alpha^{(j)}) = \prod_{k=1}^{\infty} \prod_{i=1}^{N_j} p(y_{ik}^{(j)} | \theta_k^{(j)}, \alpha_k^{(j)})$$

used by Eq. (4.1), is now replaced by:

$$\begin{aligned} p(\mathbf{Y}^{(j)} | \boldsymbol{\theta}^{(j)}, \alpha^{(j)}, \beta, \rho) &\propto \exp \left(\sum_{k=1}^{\infty} \sum_{i=1}^{N_j} \log p(y_{ik}^{(j)} | \theta_k^{(j)}, \alpha_k^{(j)}) \right. \\ &\quad \left. + \sum_{k=1}^{\infty} \log \Phi(\mathbf{Y}_{.k}^{(j)}) + \sum_{i=1}^{N_j} \log \Theta(\mathbf{Y}_{i.}^{(j)}) \right). \end{aligned}$$

and the list of model parameters Π is extended to $\Pi = \{\vartheta, \sigma_A, \sigma, \boldsymbol{\alpha}, \rho, \beta, \mathcal{M}\}$.

4.2.2.3 Comparison with Joint Topic Model

In this section, we highlight the key differences and similarities between WS-MRF-SIBP and the work of chapter 3.

1. **Finite vs. Infinite** Image prior θ is drawn from a Dirichlet distribution in the work of chapter 3. It has a fixed dimensions K , because the dimensionality k of the Dirichlet distribution is assumed known and fixed. In contrast, WS-MRF-SIBP generates θ by following the infinite stick-breaking process [189], where

4. Weakly Supervised Learning of Objects, Attributes and their Associations

each of independent random variables is drawn from Beta distribution.

2. **Topic allocation** In chapter 3, θ is the random parameter of a multinomial over topics y . This assumption does not hold in the presence of both objects and attributes. Each path is dominated by either an object or attribute topic, since object and attributes topics compete for the same patch. In contrast, WS-MRF-SIBP samples each y as an independent Bernoulli distribution given θ . The object factors and attribute factors can co-exist happily.
3. **Spatial structure** Two spatial smoothing term are specifically designed for different tasks. In chapter 3, we encode the spatial location of each observation with Normal distribution, which is to reflect the intuition that foreground objects tend to be compact. Thus, this is very efficient for localising objects in an image. In contrast, WS-MRF-SIBP integrates MRFs to enforce spatial coherence. We choose this because it can be further extended to capture the correlations of factors (e.g. objects and attributes).

4.2.3 Model Learning

Exact inference for $p(\mathbf{H}|\mathbf{X}, \Pi)$ in the proposed stacked IBP is intractable, so an approximate inference algorithm in the spirit of [189] is developed. The mean field variational approximation to the desired posterior $p(\mathbf{H}|\mathbf{X}, \Pi)$ is:

$$q(\mathbf{H}) = \prod_{j=1}^J (q_{\tau}(\mathbf{b}^{(j)})q_{\nu}(\mathbf{Y}^{(j)}))q_{\varphi}(\mathbf{A}) \quad (4.5)$$

where $q_{\tau}(b_k^{(j)}) = \text{Beta}(b_k^{(j)}; \tau_{k1}^{(j)} \tau_{k2}^{(j)})$, $q_{\nu}(y_{ik}^{(j)}) = \text{Bernoulli}(y_{ik}^{(j)}; \nu_{ik}^{(j)})$, $q_{\varphi}(\mathbf{A}_{k\cdot}) = \mathcal{N}(\mathbf{A}_{k\cdot}; \varphi_k, \varphi_k)$ and τ, ν, φ are variational parameters for b, y, \mathbf{A} respectively. The infinite stick-breaking process for latent factors is truncated at K_{max} , so $\theta_k = 0$ for $k > K_{max}$. A variational message passing (VMP) strategy [189] can be used to minimise the KL divergence of Eq. (4.5) to the true posterior. Updates are obtained by deriving integrals of the form $\ln q(\mathbf{h}) = E_{\mathbf{H} \setminus \mathbf{h}} [\ln p(\mathbf{H}, \mathbf{X})] + C$ for each group of hidden variables \mathbf{h} . These result in the series of iterative updates given in Algorithm 1, where $\Psi(\cdot)$ is the

4. Weakly Supervised Learning of Objects, Attributes and their Associations

digamma function; and $q_{ms}^{(j)}$ and $\mathbb{E}_{\mathbf{b}}[\log(1 - \prod_{t=1}^k b_t^{(j)})]$ are given in [189].

Algorithm 1: Variational Inference for WS-MRF-SIBP

```

1 while not converge do
2   for k = 1 to Kmax do
3
4       
$$\varphi_k = \left( \frac{1}{\sigma^2} \sum_{j=1}^J \sum_{i=1}^{N_j} \nu_{ik}^{(j)} (\mathbf{X}_{i\cdot}^{(j)} - \sum_{l:l \neq k} \nu_{il}^{(j)} \varphi_l) \right) \cdot \left( \frac{1}{\sigma_A^2} + \frac{1}{\sigma^2} \sum_{j=1}^J \sum_{i=1}^{N_j} \nu_{ik}^{(j)} \right)^{-1} \quad (4.6)$$

5
6       
$$\varphi_k = \left( \frac{1}{\sigma_A^2} + \frac{1}{\sigma^2} \sum_{j=1}^J \sum_{i=1}^{N_j} \nu_{ik}^{(j)} \right)^{-1} \mathbf{I} \quad \triangleright \text{Update appearance term including mean and covariance.} \quad (4.7)$$

7
8   end
9   for j = 1 to J do
10    for k = 1 to Kmax do
11
12        
$$\tau_{k1}^{(j)} = \vartheta + \sum_{m=k}^{K_{max}} \sum_{i=1}^{N_j} \nu_{im}^{(j)} + \sum_{m=k+1}^{K_{max}} (N_j - \sum_{i=1}^{N_j} \nu_{im}^{(j)}) \left( \sum_{s=k+1}^m q_{ms}^{(j)} \right) \quad (4.8)$$

13
14        
$$\tau_{k2}^{(j)} = 1 + \sum_{m=k}^{K_{max}} (N_j - \sum_{i=1}^{N_j} \nu_{im}^{(j)}) q_{mk}^{(j)} \quad \triangleright \text{Update image prior for every factor k.} \quad (4.9)$$

15
16        for i = 1 to Nj do
17
18            
$$\eta = \sum_{t=1}^k (\Psi(\tau_{t1}^{(j)}) - \Psi(\tau_{t2}^{(j)})) - \mathbb{E}_{\mathbf{b}}[\log(1 - \prod_{t=1}^k b_t^{(j)})] - \frac{1}{2\sigma^2} (\text{tr}(\varphi_k) + \varphi_k \varphi_k^T - 2\varphi_k (\mathbf{X}_{i\cdot}^{(j)} - \sum_{l:l \neq k} \nu_{il}^{(j)} \varphi_l)^T) \quad \triangleright \text{Update each latent factor} \quad (4.10)$$

19
20            
$$\eta' = \eta + \sum_{m \in N(i)} \beta \eta_{mk}^{(j)} + \sum_{n:n \neq k} \rho \mathcal{M}_{kn} \eta_{in} \quad \triangleright \text{Inferred with two types of MRF terms} \quad (4.11)$$

21
22            
$$\nu_{ik}^{(j)} = \frac{\alpha_k^{(j)}}{1 + \exp[-\eta']} \quad \triangleright \text{The final output indicate the sample state of each latent factor.} \quad (4.12)$$

23        end
24    end
25  end
26  for j = 1 to J do
27    for i = 1 to Nj do
28
29        
$$\mathcal{M} = \mathcal{M} + (\eta_{i\cdot}^j)^T \eta_{i\cdot}^j \quad \triangleright \text{The intra-superpixel correlation is disambiguated inferred factors} \quad (4.13)$$

30    end
31  end
32 end

```

Like [190, 191], the MRF influence is via Eqs. (4.11) and (4.12). However, while the work of [190, 191] only considers spatial coherence, this work further models the inter-factor correlation, which is clearly to see that it is very important for the presented WS tasks, especially in image annotation.

Factor correlation learning: The correlation matrix \mathcal{M} is non-trivial to estimate

4. Weakly Supervised Learning of Objects, Attributes and their Associations

accurately in the WSL case, in contrast to the fully-supervised case where it is easy to obtain via computing the correlation of patch-annotation. In the WSL case, it can only be estimated a priori from image-level tags. However, this is a very noisy estimate. For example, an image with tags *furry*, *horse*, *metal*, *car* will erroneously suggest *horse-car*, *furry-metal*, *horse-meta*, etc correlations.

To address this, the approach initialises \mathcal{M} coarsely with image-level labels as $\mathcal{M} = \sum_{i=j}^J (\alpha_k^{(j)})^T (\alpha_k^{(j)})$, and refine it with an EM process. During learning, the model e-estimates \mathcal{M} at each iteration using the disambiguated patch-level factors inferred by the model, as in Eq. (4.13). Thus as the correlation estimate improves, the estimated factors become more accurate, and vice-versa. The effectiveness of this iterative learning procedure is demonstrated in Section 4.3.3.

Efficiency: In practice, the truncation approximation means that the proposed WS-MRF-SIBP runs with a finite number of factors K_{max} which can be freely set so long as it is bigger than the number of factors needed by both annotations and background clutter (K_{bg}), i.e., $K_{max} \gg K_o + K_a + K_{bg}$. Despite the combinatorial nature of the object-attribute association and localisation problem, the proposed model is of complexity $\mathcal{O}(JNDK_{max} + K_{max}^2)$ for J images with N patches, D feature dimension and K_{max} truncated factors.

4.2.4 Inference for Test Data

At testing time, the appearance of each factor k , now modelled by sufficient statistics $\mathcal{N}(\mathbf{A}_k; \varphi, \boldsymbol{\varphi}_k)$, is assumed to be known (learned from the training data), while annotations for each test image $\alpha_k^{(j)}$ will need to be inferred. Thus Algorithm 1 still applies, but without the appearance update terms (Eqs. (4.6) and (4.7)) and with $\alpha_k^{(j)} = 1 \forall k$, to reflect the fact that all the learned object, attribute, and background types could be present.

4.2.5 Applications of the Model

Given the learned model applied to test data, it can perform the following tasks.

Free image annotation: This is to describe an image using a list of nouns and adjectives corresponding to objects and their associated attributes, as well as locating them. To infer what objects are present in image j , the first K_o latent factors of the

4. Weakly Supervised Learning of Objects, Attributes and their Associations

inferred $\theta^{(j)}$ are thresholded or ranked to obtain a list of objects. This is followed by locating them via searching for the superpixels i^* maximising $\mathbf{Y}_{ik}^{(j)}$, then thresholding or ranking the K_a attribute latent factors in $\mathbf{Z}_{j^*k}^{(i)}$ to describe them. This corresponds to a “*describe this image*” task.

Annotation given object names: This is a more constrained variant of the free annotation task above. Given a named (but not located) object k , its associated attributes can be estimated by first finding the location as $i^* = \arg \max_i \mathbf{Y}_{ik}^{(j)}$, then the associated attributes by $\mathbf{Y}_{i^*k}^{(j)}$ for $K_o < k \leq K_o + K_a$. This corresponds to a “*describe this (named) object in an image*” task.

Object+Attribute Query: Images can be queried for a specified object-attribute conjunction $\langle k_o, k_a \rangle$ by searching for $i^*, j^* = \arg \max_i \mathbf{Y}_{ik_o}^{(j)} \cdot \mathbf{Y}_{ik_a}^{(j)}$. This corresponds to a “*find images with a particular **kind of object***” task.

Semantic segmentation: In this application, the presented model aims to label each superpixel i with one of K_o learned object factors. The label of superpixel i can be obtained by searching $k^* = \arg \max_k \mathbf{Y}_{ik}^{(j)}$, where $k \in K_o$. Note that although the annotation search space is solely objects, inference of the additional $k > K_o$ factors (including unannotated background or attribute annotation) can help detect objects $k \in K_o$ via disambiguation. Note that unlike most WS semantic segmentation work [124, 127], the proposed model can operate with access to the whole test set. It can also operate under an the transductive setting as those models do. Under this setting, the appearance distribution $\mathcal{N}(\mathbf{A}_{k\cdot}; \varphi_k, \boldsymbol{\varphi}_k)$ will be further updated by Eqs. (4.6) and (4.7) based on the test images. The image-level label of testing data is then assigned by the inferred factors of the proposed model or alternatively by an image classifier (see Section 4.3.2.3).

4.3 Experiments

Extensive experiments were carried out to demonstrate the effectiveness of the proposed model on three real-world applications: image annotation (see Section 4.3.1.2), object-attribute query (see Section 4.3.1.3) and semantic segmentation (see Section 4.3.2).

4. Weakly Supervised Learning of Objects, Attributes and their Associations

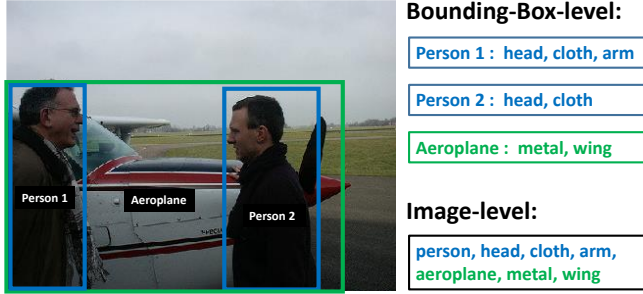


Figure 4.3: Strong bounding-box-level annotation and weak image-level annotations for aPascal are used for learning strongly supervised models and WS models respectively.



Figure 4.4: 43 subordinate classes of dog are converted into a single entry-level class ‘dog’.

4.3.1 Image Annotation and Query

4.3.1.1 Datasets and Settings

For the image annotation and query tasks, various object and attribute datasets are available such as aPascal [22], ImageNet [177], SUN [230] and AwA [36]. This work chooses aPascal because it has multiple objects per image; and ImageNet because attributes are shared widely across categories.

aPascal: This dataset [22] is an attribute labelled version of PASCAL VOC 2008. There are 4340 images of 20 object categories. Each object is annotated with a list of 64 attributes that describe them by shape (e.g., isBoxy), parts (e.g., hasHead) and material (e.g., isFurry). In the original aPascal, attributes are strongly labelled for 12695 object bounding boxes, i.e. the object-attribute association are given. To test the proposed WS approach, this work merges the object-level category annotations and attribute annotations into a single annotation vector of length 84 for the entire image. This image-level annotation is much weaker than the original bounding-box-level annotation, as shown in Figure 4.3. In all experiments, the proposed model uses the same train/test splits provided by [22].

ImageNet Attribute: This dataset [177] contains 9600 images from 384 ImageNet synsets/categories. To study WSL, this work ignores the provided bounding box annotation. Attributes for each bounding box are labelled as 1 (presence), -1 (absence) or 0 (ambiguous). This work uses the same 20 of 25 attributes as [177] and consider

4. Weakly Supervised Learning of Objects, Attributes and their Associations

1 and 0 as positive examples. Many of the 384 categories are subordinate categories, e.g. dog breeds. However, distinguishing fine-grained subordinate categories is beyond the scope of this study. That is, the presented method is interested in finding a ‘black-dog’ or ‘white-car’, rather than ‘black-mutt’ or ‘white-ford-focus’. This work thus converts the 384 ImageNet categories to 172 entry-level categories using [231] (see Figure 4.4). This work evenly splits each class to create the training and testing sets.

This work compares the proposed WS-MRF-SIBP to two strongly supervised models and four WS alternatives:

Strongly supervised models: A strongly supervised model uses bounding-box-level annotation. Two variants are considered for the two datasets respectively. **DPM+s-SVM:** for aPascal, both object detector and attribute classifier are trained from FS data (i.e. Bounding-Box-level annotation in Figure 4.3). Specifically, this work uses the 20 pre-trained DPM detectors from [16] and 64 attribute classifiers from [22]. **GT+s-SVM:** for ImageNet attributes, there is not enough data to learn 172 strong DPM detectors as in aPascal. So the ground truth bounding box is used instead assuming perfect object detectors is available, giving a significant advantage to this strongly supervised model. This work trains attribute classifiers using the presented feature and liblinear SVM [232]. These strongly supervised models are similar in spirit to the models used in [1, 138, 172] and can be considered to provide an upper bound for the performance of the WS models.

Weakly supervised models: **w-SVM** [22, 177]: In this weakly-supervised baseline, both object detectors and attribute classifiers are trained on the weak image-level labels as for the proposed model (see Figure 4.3). For aPascal, this baseline trains object and attribute classifiers using the feature extraction and model training codes (which is also based on [232]) provided by the authors of [22]. For ImageNet, the presented features are used, without segmentation. **MIML** [233]: This is the MIML learning method in [233]. In a way, the presented model can also be considered as a MIML method with each image a bag and each superpixel an instance. The MIML model provides a mechanism to use the same super-pixel/patch based representation for images as the proposed model, thus providing the object/attribute localisation capability as the proposed model does. **w-LDA:** Weakly-supervised LDA approaches [185, 188] have been used for WSOL. A generalisation of LDA [179, 188] is implemented that accepts

4. Weakly Supervised Learning of Objects, Attributes and their Associations



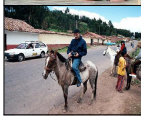
	w-SVM		MIML		w-LDA		WSDC		Ours		DPM+s-SVM	
	bicycle	motorbike	person	bicycle	motorbike	bicycle	motorbike	person	person	motorbike	person	motorbike
	metal row wind shiny text wool	metal row wind shiny text wool	skin cloth shiny leather foot/shoe	headlight window wheel arm screen	taillight engine shiny glass	wheel rein beak metal sail	metal wheel saddle furn.seat taillight	head plastic screen leg leather	cloth head nose face leaf	wheel exhaust shiny wool text	cloth skin hair head leg	engine metal label shiny taillight
	cat	sheep	cat	dog	person	cat	person	cat	person	cat	person	dog
	furry cloth snout leg head	furry cloth snout leg head	furry horn ear occluded leg	mouth furry torso taillight shiny	cloth wood mast torso arm	nose furry arm jet engine foot	skin glass face label hand	furry cloth wool head 2D Boxy	skin cloth leg arm hand rein Skin	head leg furry hand Skin	cloth torso skin red head	beak hand arm furry skin
	car	person	person	train	car	cow	horse	bus	horse	car	car	horse
	window glass pedal 3d boxy metal	window glass pedal 3d boxy metal	hand screen cloth arm flower	metal vert cyl wing door leg	wheel metal propeller label door	furn.leg head furry cloth wool	leg head metal ear hair	glass exhaust rein head 3D Boxy	furry head ear arm leg	meta window wheel 3d boxy door	metal 3d boxy plastic engine side mirror	furry torso leaf saddle feather

Figure 4.5: Qualitative results on free annotation. False positives are shown in red. If the object prediction is wrong, the corresponding attribute box is shaded.

continuous feature vectors (instead of BoW). Like MIML this method can also accept superpixel based representation, but w-LDA is more related to the proposed WS-SIBP than MIML since it is also a generative model. **WSDC** [127]: WS dual clustering is a state-of-the-art method for semantic segmentation that estimates pixel-level annotation given only image-level labels. This semantic segmentation method can be re-purposed to the presented image annotation setting by considering the same input (superpixel representation + image-level label) followed by the same method as in the proposed framework to first infer superpixel level labels and then congregate them to compute image-level annotations (see Section 4.2.5).

4.3.1.2 Image Annotation

An image description (annotation) can be automatically generated by predicting objects and their associated attributes. To comprehensively cover all aspects of performance of the proposed method and competitors, this model performs three annotation tasks with different amount of constraints on test images: (1) *free annotation*, where no constraint is given to a test image, (2) *annotation given object names*, where named but not located objects are known for each test image, and (3) *annotation given locations*, where objects locations are given in the form of bounding boxes, where the attributes can be predicted.

4. Weakly Supervised Learning of Objects, Attributes and their Associations

	aPascal [22]			ImageNet [177]		
	AP@2	AP@5	AP@8	AP@2	AP@5	AP@8
w-SVM [22]	24.8	21.2	20.3	46.3	41.1	37.5
MIML [233]	28.7	22.4	21.0	46.6	43.2	38.3
w-LDA [188]	30.7	24.0	21.5	48.4	43.1	38.4
WSDC [127]	29.8	25.1	21.3	48.0	42.7	36.5
WS-MRF-SIBP	40.1	29.7	25.0	60.7	54.2	50.0
DPM+s-SVM	40.6	30.3	23.8	65.9	60.7	53.2

Table 4.1: Free annotation performance evaluated on t attributes per object.

Free annotation: For WS-MRF-SIBP, w-LDA and MIML the procedure in Section 4.2.5 is used to detect objects and then describe them using the top t attributes. For the strongly supervised model on aPascal (DPM+s-SVM), this work uses DPM object detectors to find the most confident objects and their bounding boxes in each test image. Then the 64 attribute classifiers is used to predict top t attributes in each bounding box. In contrast, w-SVM trains attributes and objects independently, and cannot associate objects and attributes. This work thus uses it to predict only one attribute vector per image regardless of which object label it predicts.

Since there are variable number of objects per image in aPascal, quantitatively evaluating free annotation is not straightforward. Therefore, this work evaluates only the most confident object and its associated top t attributes in each image, although more could be described. For ImageNet, there is only one object per image. This work follows [234, 235] in evaluating annotation accuracy by Average Precision (AP), given varying numbers (t) of predicted attributes per object. Note that if the predicted object is wrong, all associated attributes are considered wrong.

Table 4.1 compares the free annotation performance of the compared models. This work has the following observations: (1) the proposed WS-MRF-SIBP, despite learned with the weak image-level annotation, yields comparable performance to the strongly supervised model (DPM/GT+s-SVM). The gap is particularly small for the more challenging aPascal dataset, whilst for ImageNet, the gap is bigger as the strongly supervised GT+s-SVM has an unfair advantage by using the ground truth bounding boxes during testing. (2) WS-MRF-SIBP consistently outperforms the four WS alternatives. The margin is especially large for $t = 2$ attributes per object, which is closest to the

4. Weakly Supervised Learning of Objects, Attributes and their Associations

true number of attributes per object. For bigger t , all models must generate some irrelevant attributes thus narrowing the gaps. (3) As expected, the w-SVM model obtains the weakest results, suggesting that the ability to locate objects is important for modelling object-attribute association. (4) Compared to the two generative models (ours and w-LDA), MIML has worse performance because a generative model is more capable of utilising weak labels [188]. The other discriminative model WSDC fares better than MIML due to the ability to exploit superpixel appearance similarity to disambiguate the image-level labels, but it is still much inferior to the proposed model. (5) Between the two generative models, the advantage of the proposed framework over w-LDA is clear; due to the ability of IBP to explain each superpixel with multiple non-competing factors¹.

Figure 4.5 shows some qualitative results on aPascal via the two most confident objects and their associated attributes. This is a challenging dataset – even the strongly supervised DPM+s-SVM makes mistakes for both attribute and object prediction. Compared to the WS models, WS-MRF-SIBP has more accurate prediction – it jointly and non-competitively models objects and their attributes so object detection benefits from attribute detection and vice versa. Other WS models are also more likely to mismatch attributes with objects, e.g. MIML detects a shiny person rather than the correct shiny motorbike.

To gain some insight into what has been learned by the proposed model and why it is better than the WS alternatives, Figure 4.6 visualises the attribute and object factors learned by WS-SIBP model and MIML and w-LDA which also use superpixels as input. It is evident that without explicit background modelling, MIML suffers greatly by trying to explain the background superpixel using the weak labels. In contrast, both w-LDA and WS-SIBP have good segmentation of foreground objects, showing that both the learned foreground and background topics are meaningful. However, for w-LDA, since object and attributes topics compete for the same superpixel, each superpixel is dominated by either an object or attribute topic. In contrast, the object factors and attribute factors co-exist happily in WS-SIBP as they should do, e.g. most person superpixels have the clothing attribute as well.

Annotation given object names (GN): In this experiment, it assumes that object la-

¹Training two independent w-LDA models for objects and attributes respectively is not a solution: the problem would re-occur for multiple competing attributes.

4. Weakly Supervised Learning of Objects, Attributes and their Associations

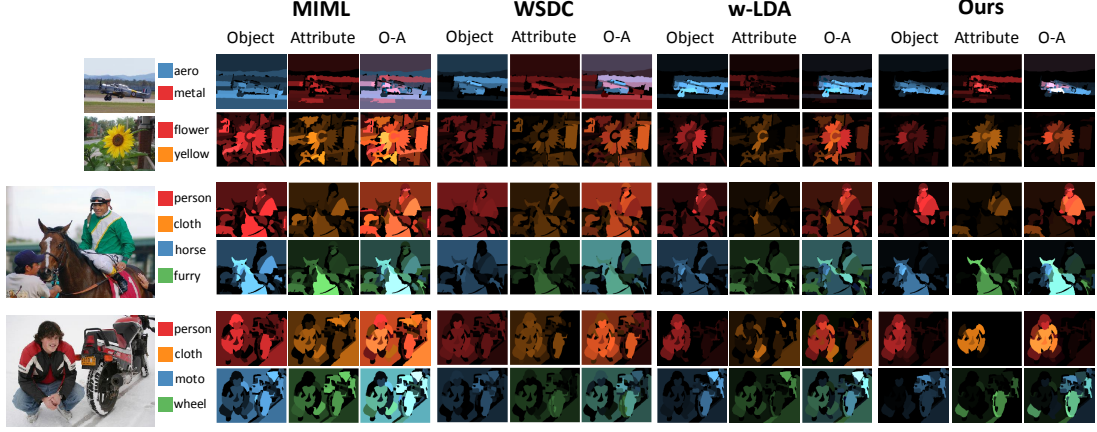


Figure 4.6: Illustrating the inferred patch-annotation. Object and attributes are coloured, and multi-label annotation blends colours. The bottom two groups each have two rows corresponding to the two most confident objects detected.

bels are given and the approaches aim to describe each object by attributes, corresponding to tasks such as: “Describe the car in this image”. For the strongly supervised model on aPascal, the object’s DPM detector is used to find the most confident bounding box. Then attributes are predicted for that box. Here, annotation accuracy is the same as attribute accuracy, so the performance of different models is evaluated following [236] by Mean Average Precision (MAP) under the Precision Recall (PR) curve. Note that for aPascal, w-SVM reports the same list of attributes for all co-existing objects, without being able to localise and distinguish them. Its result is thus not meaningful and is excluded. The same set of conclusions can be drawn from Table 4.2 as in the free annotation task: the proposed WS-MRF-SIBP at par with the supervised models and outperforming the WS ones.

		w-SVM	MIML	w-LDA	WSDC	Ours	SS
GN	aPascal	—	32.1	35.5	36.3	39.3	41.8
	ImageNet	32.4	33.5	39.6	44.2	52.8	56.8
GL	aPascal	33.2	35.1	35.8	38.4	43.6	42.1
	ImageNet	37.7	39.1	46.8	48.2	53.9	56.8

Table 4.2: Results on annotation given object names (GN) or locations (GL). SS stands for Strongly Supervised.

4. Weakly Supervised Learning of Objects, Attributes and their Associations

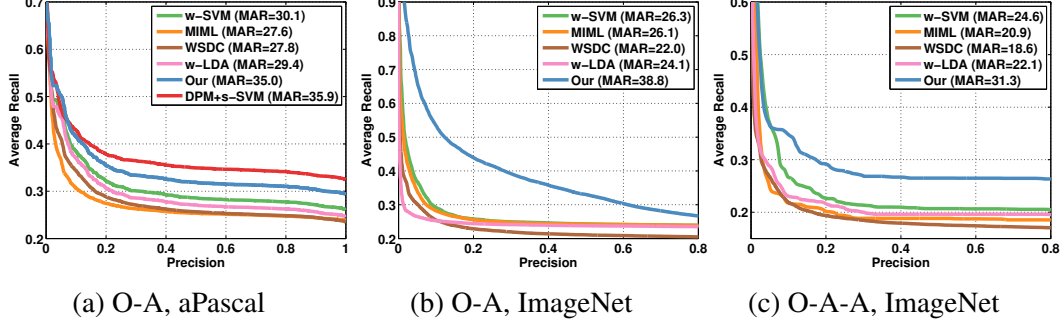


Figure 4.7: Object-attribute query results as precision-average recall curve.

Given object location (GL): If the bounding box of an object is further known in a test image, it can simply predict attributes inside each bounding box. This becomes the conventional attribute prediction task [22, 177] for describing an object. Table 4.2 shows the results, where similar observations can be made as in the other two tasks above. Note that in this case the strongly supervised model is the method used in [22]. The mAP obtained using the proposed WS model is even higher than the strongly supervised model (though the presented area-under-ROC-curve value of 81.5 is slightly lower than the 83.4 reported in [22]).

4.3.1.3 Object-attribute Query

In this task object-attribute association is used for image retrieval. Following work on multi-attribute queries [27], this section uses Mean Average Recall (MAR) over all precisions (MAR) as the evaluation metric. Note that unlike [27] which requires each queried *combination* to have enough (100) training examples to train conjunction classifiers, the proposed method can query novel never-previously-seen combinations. Three experiments are conducted. This experiment generates 300 random object-attribute combinations for aPascal and ImageNet respectively and 300 object-attribute-attribute queries for ImageNet. For the strongly supervised model, the model normalises and multiplies object detector with attribute classifier scores. No object detector is trained for ImageNet so no result is reported there. For w-SVM, this experiment uses [176] to calibrate the SVM scores for objects and attributes as in [27]. For the three WS models, the procedure in Section 4.2.5 is used to compute the retrieval

4. Weakly Supervised Learning of Objects, Attributes and their Associations

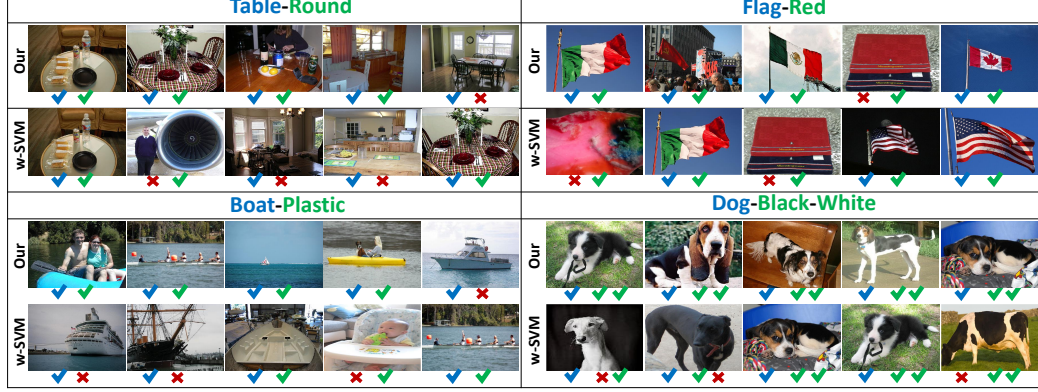


Figure 4.8: Object-attribute query: qualitative comparison

ranking.

Quantitative results are shown in Figure 4.7 and some qualitative examples in Figure 4.8. The proposed WS-SIBP has a very similar MAR values to the strongly supervised DPM+s-SVM, while outperforming all the other models. w-SVM calibration [176] helps it outperform MIML and w-LDA. However, the lack of object-attribute association and background modelling still causes problems for w-SVM. This is illustrated in the ‘dog-black-white’ example shown in Figure 4.8 where a white background caused an image with a black dog retrieved at Rank 2 by w-SVM.

4.3.2 Semantic Segmentation

4.3.2.1 Datasets and Settings

This section evaluates the semantic segmentation performance of the presented model on aPascal Segmentation dataset (Section 4.3.2.2) and LabelMe Outdoor dataset (Section 4.3.2.3) under the WS setting.

aPascal Segmentation: This dataset [123] is a subset of PASCAL VOC 2008 [22] where both pixel-level segmentation and attributes annotation are available. It contains 639 images from 20 classes. The 64 attribute annotation for each image is same with the aPascal dataset used in the annotation experiments. This section uses the training (326 images) and testing (313 images) split provided by [123].

LabelMe Outdoor Segmentation: Also knowns as the SIFT Flow dataset [117], this widely used dataset contains 2688 images densely labelled with 33 object classes at pixel-level using the LabelMe online annotation tool. Every pixel in each image is

4. Weakly Supervised Learning of Objects, Attributes and their Associations

assigned a label meaning that background ‘stuffs’ such as sky, sea, street are also labelled as objects. Most images contain outdoor scenes. This experiment uses the standard training (2488 images) and testing (200 images) split provided in [117]. Note that no attribute label is available for this dataset.

Evaluation metrics: The evaluation metrics used for semantic segmentation are often dataset dependent. Past results on the LabelMe dataset typically report results in both total per-pixel accuracy, which measures the percentage of correctly labelled pixels in the test images, and per-class accuracy, which is the percentage of correctly labelled pixels for a class and then averaged over all object classes. Both metrics are necessary because for any model, some model parameters can typically be tuned so that one metric gets higher at the price of lowering the other metric. The VOC images have very different characteristics compared with LabelMe. In particular, the images often contain large portions of unannotated background (stuffs) and the 20 objects of interest are relatively small. The Intersection-over-union (IOU) score is thus typically used for semantic segmentation performance evaluation on the Pascal VOC dataset [123, 127, 237] and adopted here on aPascal.

4.3.2.2 Results on aPascal

To the best of our knowledge, no previous work models objects and attributes jointly for semantic segmentation under the WS setting. Note the methods proposed in [35] and [238] aim to discover and recognize basic entities from image collections in a fully unsupervised manner. The key insight to their methods is to discover the objects present in the images by analysing unlabelled data and searching for re-occurring patterns. Differently, this work focuses on segmenting out specific objects under the weakly supervised setting.

This work therefore applies the state-of-the-art WS (object only) segmentation method WSDC [127] as an alternative (see Section 4.3.1.1). With the proposed WS-MRF-SIBP, the association of objects and available attributes can be leveraged to improve performance. the presented method explores three different potential sets of attribute annotations: (1) 8 attributes: the 8 material-type of attributes selected by [123] for joint object-attribute segmentation. (2) 64 attributes: the original attributes provided by [22]. (3) 74 attributes: this experiment adds 10 more color attributes based

4. Weakly Supervised Learning of Objects, Attributes and their Associations

on aPascal sentence descriptions [1, 239].

Method		Avg. IOU (%)
Str.	Zheng <i>et al.</i> [123]	37.1
	Krähenbühl <i>et al.</i> [240]	36.9
Weak	WSDC [127]	18.2
	Ours_0attribute	23.6
	Ours_8attributes	27.3
	Ours_64attributes	28.9
	Ours_74attributes	29.4

Table 4.3: Quantitative semantic segmentation comparison versus state-of-the-art on the aPascal dataset.

The performance of the proposed model is compared with one WS [127] and two fully-supervised alternative models [123, 240] in Table 4.3. It can be observed that the proposed method outperforms the alternative weakly-supervised model WSDC [127], even without attribute annotation. Moreover, it gradually improves as more attribute annotation becomes available, and eventually its performance with 74 attributes is not far off compared to the two fully-supervised models. Note that Li *et al.* [129] shares a similar spirit to ours in the sense of exploiting attributes relations. However, they focus on the task of segmentation-based classification, which makes it difficult to directly compare with our segmentation results.

Some qualitative results are depicted in Figure 4.9. In each example, this section shows the color-coded segmentation output of Ours_0attribute, Ours_64attribute and ground truth segmentation. Coloured regions are identified as the same foreground objects. It can be noticed that large intra-class appearance variability can confuse Ours_0attribute as shown in the horse example (the third row of Figure 4.9). However, by exploiting the additional (weak) attribute annotation, the segmentation performance is greatly improved (Ours_64attribute) through disambiguation, and by capturing object-attribute co-occurrence.

4. Weakly Supervised Learning of Objects, Attributes and their Associations



Figure 4.9: Qualitative illustration of (attribute-enhanced) semantic segmentation results on aPascal.

4. Weakly Supervised Learning of Objects, Attributes and their Associations

4.3.2.3 Results on LabelMe

Method		Per-pixel (%)	Per-class (%)
Strong	Tighe <i>et al.</i> [241]	77.0	30.1
	Tighe <i>et al.</i> [122]	78.6	39.2
	Sigh and Kosecka [118]	79.2	33.8
	Yang <i>et al.</i> [121]	79.8	48.7
	Gould <i>et al.</i> [242]	78.4	25.7
Weak	Vezhnevets <i>et al.</i> [243]	-	14
	Vezhnevets <i>et al.</i> [125]	-	21
	Xu <i>et al.</i> [124]	21.9	27.9
	WSDC [127]	19.3	25.0
	Ours	46.2	23.8
	Ours_predict	48.1	26.7
	Ours_transductive	52.5	31.2

Table 4.4: Quantitative comparison of semantic segmentation performance on the LabelMe dataset.

Table 4.4 compares the performance of the proposed model with a number of state-of-the-art FS [118, 121, 122, 241, 242] and WS [124, 125, 127] models. Three variants of the proposed models has been evaluated: Ours and Ours.transductive differ in whether the test set images are used for model update (see Section 4.2.5), whilst for Ours_predict, this experiment follows [124] and use a pre-trained multi-label image classifier (Linear SVM with ImageNet-trained CNN features as input) to predict image-level object labels and use the labels for transductive learning.

The results show that the proposed model outperforms the alternative WS models in [124, 125, 127] by a large margin in both per-pixel and per-class accuracy but particularly in the per-pixel accuracy which reflects more on the performance on the large classes such as sky and sea. Note that most WSL methods are transductive. But the proposed model, even without accessing the whole test set (Ours), can beat the alternative WS models. When the proposed model operates in the transductive mode (Ours.transductive), the margin over the other transductive models including [124, 127] gets even bigger. It is worth mentioning that the result of Xu *et al.* [124] is obtained using the predicted image-level labels for transductive learning on the test

4. Weakly Supervised Learning of Objects, Attributes and their Associations

set. The presented result (Ours_transductive vs. Ours_predictive) suggests that this additional step is not necessary using the proposed model – as demonstrated in the image annotation experiments earlier, the proposed model itself can predict image-level labels and does not require assistance from another model. Table 4.4 also shows the performance of a number of state-of-the-art strongly supervised learning models [118, 121, 122, 241, 242] which require pixel-level annotation of the training images. As can be seen, there is still a fairly big gap between the best result achieved by the proposed model (Ours_transductive) and theirs, although on the per-class metric, it is close. The main reason is that in LabelMe, background stuffs such as sky and road are labelled as objects. This limits the scope for the proposed WS-MRF-SIBP to learn latent factors for the background as in the aPascal dataset, thus losing some of the power for disambiguating the weak labels.

4. Weakly Supervised Learning of Objects, Attributes and their Associations

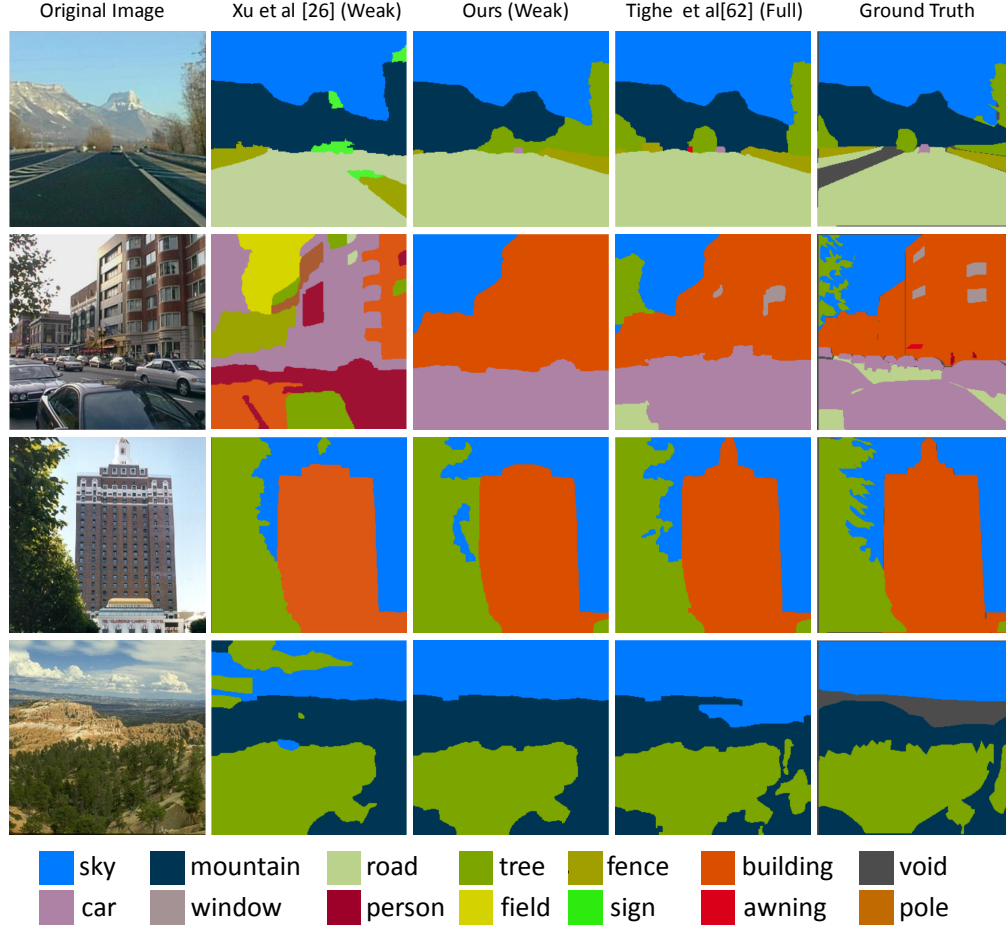


Figure 4.10: Qualitative comparison of the proposed semantic segmentation versus alternatives on the LabelMe dataset.

Figure 4.10 qualitatively compares two WS methods ([124] and Ours) and one FS method ([241]). It is noted that the advantage of the proposed model over that in [124] is particularly pronounced in the cluttered street scene (second row of Figure 4.10). For a scene like this, the ability of inferring latent factors which correspond to latent attributes for describing object appearance is critical. For example, the proposed model seems to be able to capture the fact that sky can have different types of appearance: clear and blue in the first row, overcast in the second, and cloudy in the bottom. Without accounting for these variations of appearance, the model in [124] struggled and assigned wrong labels to sky in the second and bottom row images.

4.3.3 Further Evaluations

	Free annotation		Segmentation	
	aPascal	ImageNet	LabelMe	aPascal
WS-SIBP	38.6	58.5	22.4	16.6
WS-SIBP+MRF1	38.8	58.6	43.1	22.8
WS-SIBP+MRF2	39.2	59.2	28.4	19.5
WS-SIBP+MRF2 ⁺	39.7	60.3	32.6	20.1
WS-MRF-SIBP	40.1	60.7	46.2	27.3

Table 4.5: Evaluation of individual components of the proposed model. This table reports AP@2 for free annotation on aPascal and ImageNet dataset. For segmentation, per-pixel results are reported for LabelMe and IOU for aPascal. For segmentation result on aPascal 8 attribute annotations are used.

Contributions of individual model components: To evaluate the impact of each component of the proposed model, the presented method performs experiments with several stripped-down versions of the proposed model: a) **WS-SIBP**: The basic model without any MRF assumption. b) **WS-SIBP+MRF1**: This enables the spatial MRF to smooth neighbouring superpixels. c) **WS-SIBP+MRF2**: This enables the factorial MRF to model the correlation of factors, but using only the initial image label prior. d) **WS-SIBP+MRF2⁺**: Same as c) except the correlation matrix is further updated with Eq. (4.13). e) **WS-MRF-SIBP**: The presented full model with all components (b+d).

The results in Table 4.5 show that each component contributes to the good performance of the presented method. As expected, the factorial MRF is more helpful (38.6→39.7, 58.5→60.3) for free annotation compared to spatial MRF (38.6→38.8, 58.5→58.6). In contrast, the spatial MRF significantly improves the semantic segmentation performance (22.4→43.1, 17.9→24.5) while less improvement on performance is obtained by factorial MRF (22.4→32.6, 17.9→21.8). This is expected as the spatial smoothness of the superpixel labels is perhaps the strongest cue exploited by all existing segmentation models.

Iterative learning of factorial correlation: As mentioned in Section 4.2.3, the within-superpixel factorial correlation matrix \mathcal{M} is initialised using image-level co-occurrence statistics and iteratively estimated at the superpixel-level using an EM algorithm. This process is illustrated in Figure ?? for the aPascal dataset, where the initial coarsely

4. Weakly Supervised Learning of Objects, Attributes and their Associations

estimated correlation (Figure 4.11a) is refined (Figure 4.11b) to become cleaner and closer to the ground-truth pixel-level correlation (Figure 4.11c).

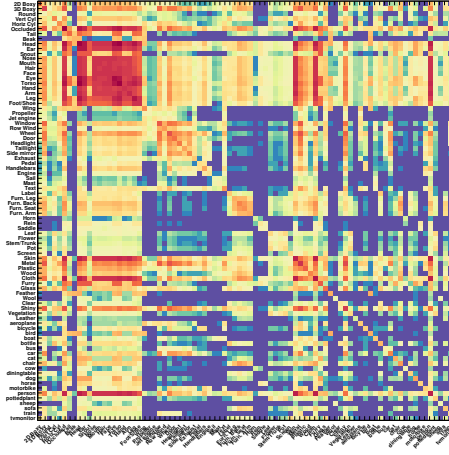
Running cost: The unoptimised single-core implementation of the proposed model requires around 2 seconds for 100 images per iteration on a PC with Intel 3.47 GHz CPU and 16GB RAM. The running time is also affected by the number of segments and latent factors K_{max} .

4.4 Summary

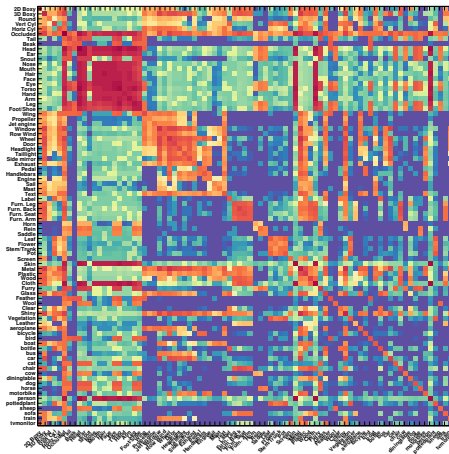
This chapter has presented an effective model for WSL of objects, attributes, their location and associations. Learning object-attribute association from weak supervision is non-trivial but critical for learning from natural data, and scaling to many classes and attributes. This work achieves this for the first time through a novel weakly-supervised stacked IBP model that simultaneously disambiguates patch-annotation correspondence, as well as learning the appearance of each annotation. The results show that the proposed model performs comparably with a strongly supervised alternative that is significantly more costly to supervise.

Nevertheless, another notable fact is that there are numerous domains in the existing natural datasets, where some domains are irrelevant and others are closely related. Obviously, transferring knowledge from a source domain is very useful for discovering target domain. In addition, it can help to reduce the amount of manual annotation by utilising the existing labels from a source domain. All these information is not considered in this chapter. To overcome these limitations, the following chapter extends the current model to a transfer learning framework. The proposed model is trained on an auxiliary dataset, which are either weakly or strongly labelled. The learned model can then be transferred and adapted to provide a powerful semantic description for various vision applications.

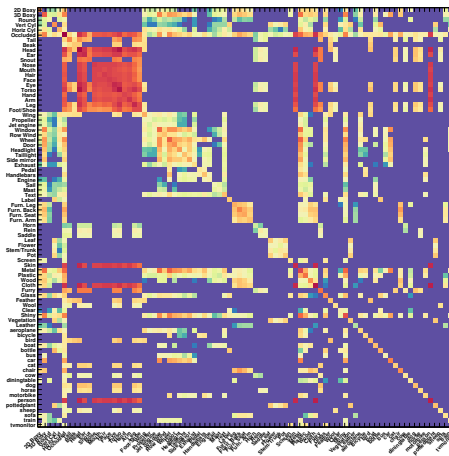
4. Weakly Supervised Learning of Objects, Attributes and their Associations



(a) Initial estimation



(b) The proposed model's refinement



(c) Ground-truth

Figure 4.11: Object-attribute query results as precision-average recall curve.

Chapter 5

Transferring a Semantic Representation for Person Re-identification and Search

5.1 Overview

The preceding chapter has demonstrated the effectiveness of the proposed model for the understanding of objects, attributes and their associations from weakly labelled data. Nevertheless, the proposed method only focuses on the target domain data. There are still some limitations that need to be solved: a) It is worth to note that enormous amounts of visual data are easily available today, especially for WS data. These images are collected from different domains, such as high quality images from a photographer or noisy images surveillance cameras. It is interesting to see that if we can transfer knowledge from a source domain to a related target domain. This is how human visual system understands a new image with learned knowledge. b) Transferring knowledge from an existing domain provides a new perspective to address the WSL problem. It can further reduce the manual labelling effort, without requiring any annotations of the target domain data. This is especially true when we learned a reliable model using an auxiliary dataset. c) Another advantage of transferring knowledge from a source domain is that it can further exploit the rich information from a fully labelled source, which can provide more explicit and reliable cues.

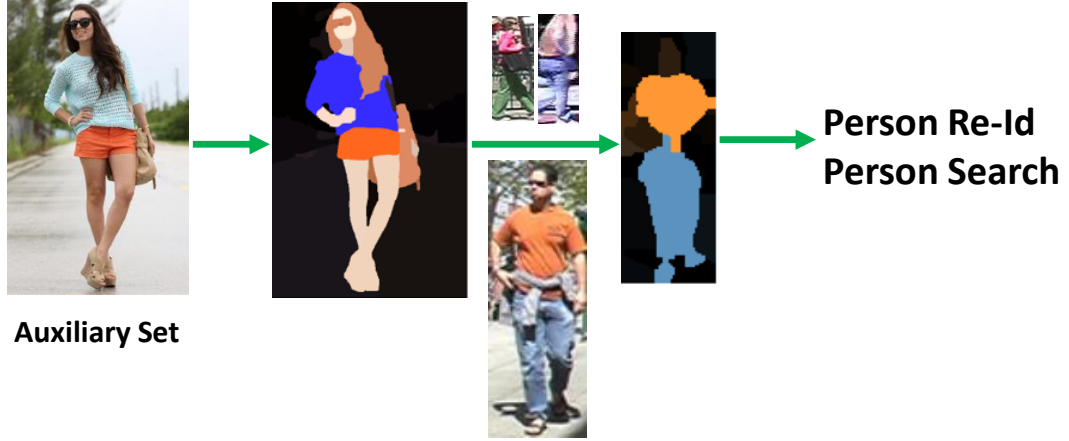


Figure 5.1: Transferring knowledge from fashion data to surveillance

In this chapter, we extended the model proposed in chapter 4 to a transfer learning framework. This framework aims to learn a semantic representation from an auxiliary dataset for various vision applications (see Figure 5.1). Specifically, this chapter generalises the proposed model to learn a semantic attribute representation. The model is trained on existing fashion photography datasets either weakly or strongly labelled. It can then be transferred and adapted to provide a powerful semantic description of surveillance person detections, without requiring any surveillance domain supervision. The resulting representation is useful for both unsupervised and supervised person Re-ID, achieving state-of-the-art and near state-of-the-art performance respectively. Furthermore, as a semantic representation it allows description-based person search to be integrated within the same framework (see Figure 5.2).

Learning semantic attributes for person Re-ID and description-based person search has gained increasing interest due to attributes great potential as a pose and view-invariant representation. However, existing attributecentric approaches have thus far underperformed state-of-the-art conventional approaches. This is due to their nonscalable need for an extensive domain (camera) specific annotation. In contrast to most existing approaches to attribute detection [57, 58] which are based on discriminative modelling, we take a generative modelling approach based on the IBP[49]. The generative formulation provides key advantages including: joint learning of all attributes; ability to naturally exploit weakly-annotated (image-level) training data; as well as unsupervised domain adaptation through Bayesian priors. Importantly an IBP-based

model [59, 60, 61] provides the favourable property of combining attributes factorially in each local patch. This means that the proposed model can differentiate potentially ambiguous situations such as Red-Shirt+Blue-Jeans versus Red-Jeans+Blue-Shirt (See Figure 5.5). Moreover, with this representation, attribute combinations that were rare or unseen at training time can be recognised at test time so long as they are individually known (e.g. Shiny-Yellow-Jeans).

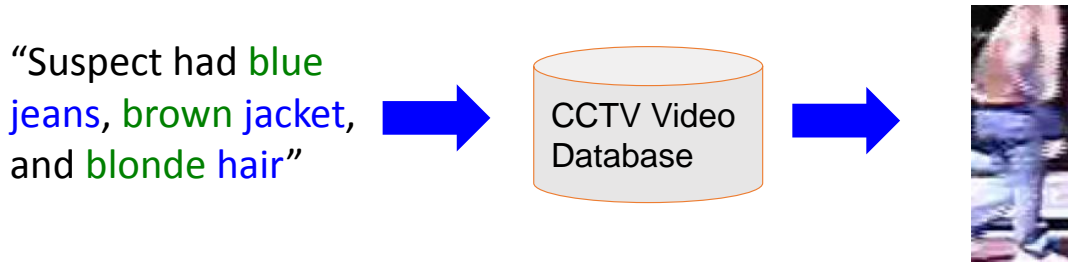


Figure 5.2: Illustration of surveillance person search procedure

The remainder of this chapter is organised as follows: Section 5.2 presents the proposed semantic representation learning framework. This includes the explanation of model learning and inference for test data. The various applications of the learned model are also discussed in Section 5.3. Experimental results are reported in Section 5.4. Finally, a summary is given in Section 5.5.

5.2 Semantic Representation Learning

The proposed model, generalised from the previous proposed WS-MRF-IBP, works with super-pixel segmented person images. Each super-pixel is associated with a K dimensional latent binary vector whose elements (factors) indicate what set of attribute properties are possessed by that patch, and which of a potentially infinite set of background clutter types are present. The first K_s factors are associated with known annotations, while the subsequent entries are free to model unannotated aspects of the images. The presented pipeline exploits two datasets: an annotated auxiliary dataset, and an unannotated target dataset (see Figure 5.3):

Auxiliary Training: First this model trains on the auxiliary/source dataset using weak (image-level) or strong (patch-level) supervision. The supervision is a binary vector

5. Transferring a Semantic Representation

describing which attributes appear in the image/patch. For strong supervision annotations $\alpha_i^{(j)}$ are given for the first K_s factors of patches i in image j , and the model learns from this each of the K_s factors' appearance. The K_s supervised factors also include foreground versus Background Patch (BGP) annotation. For weak supervision, annotations $\alpha^{(j)}$ are given for the first K_s factors of each image, and the model solves the (more challenging task) of learning both factor appearance and infer which image-level factors occur in which patches.

Target Adaptation: This model then uses the learned parameters from the auxiliary set as a prior, and adapt it to the target dataset without any supervision. The representation learned here is then used for Re-ID and person-search.

5.2.1 Model Formulation

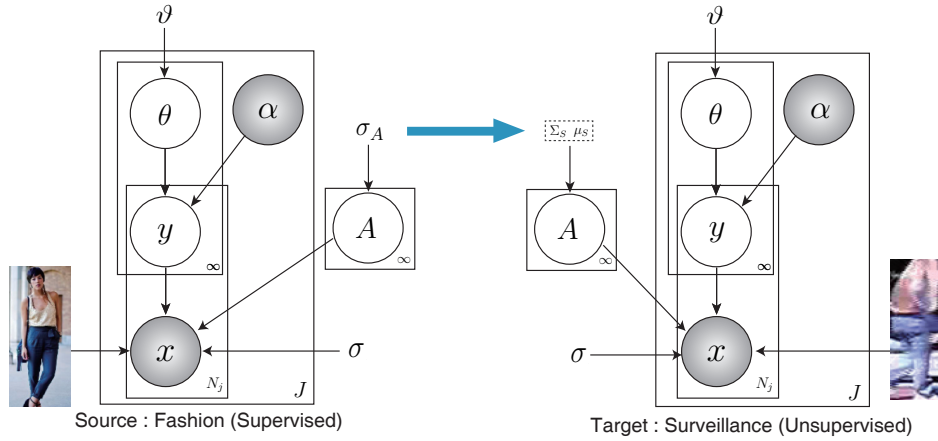


Figure 5.3: The transfer learning framework.

Each image j is represented as a bag of patches $\mathbf{X}^{(j)} = \{\mathbf{X}_i^{(j)}\}$, where \mathbf{X}_i means the vector of row i in matrix \mathbf{X} and corresponds to a D -dimensional feature representation of each patch. Without supervision, the generative process for each image is as follows:

For each latent factor $k \in 1 \dots \infty$:

1. Draw an appearance distribution mean $\mathbf{A}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

For each image $j \in 1 \dots J$:

5. Transferring a Semantic Representation

1. Draw the binary indicator matrix \mathbf{Z} describing the factor activation for every patch:

$$\begin{aligned}
 p(\mathbf{Y}^{(j)}|\vartheta, \beta) &\propto \frac{\vartheta^{K_+}}{\prod_{i=1}^{N_j} K_1^i!} \cdot \exp -\vartheta \sum_{i=1}^{N_j} \frac{1}{i} \\
 &\cdot \prod_{k=1}^{K_+} \frac{(N_j - m_k^{(j)})!(m_k^{(j)} - 1)!}{N_j!} \\
 &\cdot \prod_{k=1}^{K_+} \exp \left(\sum_{i=1}^{N_j} \beta \sum_{i' \in N(i)} \mathbf{I}(y_{ik}^{(j)} = y_{i'k}^{(j)}) \right) \quad (5.1)
 \end{aligned}$$

2. For each super-pixel patch $i \in 1 \dots N_j$: Sample patch appearance: $\mathbf{X}_i^{(j)} \sim \mathcal{N}(\mathbf{Y}_i^{(j)} \mathbf{A}, \sigma_X^2 \mathbf{I})$.

Notations: \mathcal{N} corresponds to Normal distribution with the specified parameters; ϑ is the prior expected sparsity of annotations; β is the coupling strength of the inter-patch MRF; μ_k and σ_A are the prior mean and covariance of each factor. $p(\mathbf{Y}^{(j)})$ in Eq. (5.1) corresponds to the prior of the proposed framework. It expresses the IBP sampling of an unbounded number of factors in each patch (first two lines), which are spatially correlated across patches by a Potts model MRF (last line) like [190]. Here $K_+ \geq K_s$ refers to the (inferred) number of active factors in the image, K_1^i is the factor history [49], and $m_k^{(j)}$ is the number of times each factor k is active.

Denote the set of hidden variables by $\mathbf{H} = \{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(J)}, \mathbf{A}\}$, observed images by $\mathbf{X} = \{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(J)}\}$, and model parameters by $\Pi = \{\vartheta, \beta, \sigma_X, \Sigma_k, \mu_k\}$. Then the joint probability of the variables and data given the parameters is:

$$\begin{aligned}
 p(\mathbf{H}, \mathbf{X}|\Pi) &= \prod_{k=1}^{\infty} p(\mathbf{A}_k | \mu_k, \Sigma_k) \\
 &\cdot \prod_{j=1}^J p(\mathbf{Y}^{(j)}; \vartheta, \beta) \prod_{i=1}^{N_j} p(\mathbf{X}_i^{(j)} | \mathbf{Y}_i^{(j)}, \mathbf{A}, \sigma_X) \quad (5.2)
 \end{aligned}$$

Learning in the proposed model aims to compute the posterior $p(\mathbf{H}|\mathbf{X}, \Pi)$ for: discovering which factors (object/attributes) are active on each patch (inferring $\mathbf{Y}^{(j)}$), and learning the appearance of each factor (inferring \mathbf{A}_k).

5.2.2 Model Learning from the Auxiliary Set

To learn the proposed model, this chapter exploits Gibbs sampling for approximate inference inspired by [192]. For Gibbs sampling, it is needed to derive an update for each hidden variable conditional on the observations and all the other hidden variables.

Unsupervised Factor Updates: For all initialised factors k , the presented method can sample the state of each latent factor $y_{ik}^{(j)}$ via:

$$\begin{aligned} p(y_{ik}^{(j)} = 1 | \mathbf{Y}_{-ik}^{(j)}, \mathbf{X}_i^{(j)}) &\propto p(\mathbf{X}_i^{(j)} | \mathbf{Y}^{(j)}) P(y_{ik}^{(j)} = 1 | \mathbf{Y}_{-ik}^{(j)}) \\ &= \frac{m_k^{(j)} - y_{ik}^{(j)}}{N_j} \cdot \exp \sum_{i' \in N(i)} \beta \mathbf{I}(y_{ik}^{(j)} = y_{i'k}^{(j)}) \\ &\cdot \exp\left(-\frac{1}{2\sigma_X^2} \text{tr}(\mathbf{X}_i^{(j)} - \mathbf{Y}_i^{(j)} \mathbf{A})^T (\mathbf{X}_i^{(j)} - \mathbf{Y}_i^{(j)} \mathbf{A})\right) \end{aligned} \quad (5.3)$$

where $\mathbf{Y}_{-ik}^{(j)}$ denotes the entries of $\mathbf{Y}^{(j)}$ other than $\mathbf{Y}_{ik}^{(j)}$. To sample new latent factors, $\text{Poisson}(\frac{\vartheta}{N_j})$ is used as the expected number of new classes [49].

Supervised Factor Updates: Eq. (5.3) describes inference in the case where no supervision is available. If strong supervision $\alpha_{ik}^{(j)}$ is available, Eq. (5.3) is replaced with $y_{ik}^{(j)} = \alpha_{ik}^{(j)}$. If weak supervision $\alpha_k^{(j)}$ is available Eq. (5.3) is replaced with $p(y_{ik}^{(j)} = 1 | \mathbf{Y}_{-ik}^{(j)}, \mathbf{X}_i^{(j)}) \propto p(\mathbf{X}_i^{(j)} | \mathbf{Y}^{(j)}) P(y_{ik}^{(j)} = 1 | \mathbf{Y}_{-ik}^{(j)}) \cdot \alpha_k^{(j)}$

Appearance Updates: In order to sample factor appearance \mathbf{A} , its Gaussian posterior $p(\mathbf{A} | \mathbf{X}, \mathbf{Y})$ is computed:

$$\begin{aligned} \mu_S &= (\tilde{\mathbf{Y}}^T \tilde{\mathbf{Y}} + \frac{\sigma_X^2}{\sigma_A^2} I)^{-1} \tilde{\mathbf{Y}}^T \tilde{\mathbf{X}} \\ \Sigma_S &= \sigma_X^2 (\tilde{\mathbf{Y}}^T \tilde{\mathbf{Y}} + \frac{\sigma_X^2}{\sigma_A^2} I)^{-1} \end{aligned} \quad (5.4)$$

where $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$ are the matrices that vertically concatenate the factor matrix and patch feature matrix for all images. Here μ_S is the $K_+ \times D$ matrix of appearance for each factor, and Σ_S is the $K_+ \times K_+$ matrix of variance parameters for each factor appearance. Since this is the auxiliary set, this thesis has assumed an uninformative prior, i.e., $\mathbf{A}_k \sim \mathcal{N}(0, \sigma_A)$.

5.2.3 Model Adaptation to the Target Set

In the target set, there is no supervision, so Eq. (5.3) is used to update the latent factors. The appearance updates however are changed to reflect the top-down influence from the learned auxiliary domain. Thus the target appearance μ_T is updated using the sufficient statistics from the source Σ_S and μ_S (Eq. (5.4)) as the prior:

$$\begin{aligned}\mu_T &= \Sigma_T(\sigma_X^{-2}\tilde{\mathbf{Y}}^T\tilde{\mathbf{X}} + \Sigma_S^{-1}\mu_S) \\ \Sigma_T &= \sigma_X^2(\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} + \sigma_X^2\Sigma_S^{-1})^{-1}\end{aligned}\tag{5.5}$$

5.3 Semantic Representation Applications

After the semantic representation learning described previously, each target image i is now described by a binary factor matrix \mathbf{Y}^j containing the inferred factor vector $\{\mathbf{y}_i\}_{i=1}^{N_j}$ for each superpixel i . This representation could be used directly, but find it convenient to convert it into a fixed-size representation per image. This model thus generates multiple heat-maps T_k per image representing the k th factor activation. Similar to [151], the presented method divides each image into 14 overlapping patches sampled on a 2×7 regular grid with a 32×32 window.¹ Each grid-patch is now represented by a K -dimensional attribute vector obtained by summing \mathbf{y}_i for every pixel.

5.3.1 Person Re-identification

The proposed semantic person representation can be used for both unsupervised or supervised Re-ID, according to whether a matching model is learned from the identity annotation of a given person Re-ID dataset.

Unsupervised Matching: Each image is represented by 14 patches each with a K dimensional descriptor. The person match is now converted to a semantic patch matching problem. This method adopts a patch matching algorithm TreeCANN [244] to efficiently compute the distance between images.

¹Note that the overlapping area between two neighbouring patches depends on the size of the image.

Supervised Matching: The 14 patch descriptors are concatenated to obtain an image-level descriptor. This is used as input to a recent metric learning algorithm kLFDA [151].

5.3.2 Person Search

Recall that K heat maps T_k are generated from the K factors. The probability of factor k appearing in an image can be obtained by $\max(T_k)$. When querying two or more factors, there are two possible query semantics: To query the probability of two factors both appearing anywhere in an image without preference for co-location (e.g., Coat + Bag), this method uses $\max(T_k) \cdot \max(T_{k'})$. In contrast, to query two factors that should *simultaneously* appear in the same place (e.g., Blue-Jeans), $\max(T_k \cdot T_{k'})$ is used.

5.4 Experiments

5.4.1 Datasets and Settings

Auxiliary Datasets: Two datasets are used as auxiliary sources. **Colourful-Fashion** [52] includes 2682 images. Pixel-level annotation is provided with 13 colour labels (e.g., brown, red) and 23 category labels (e.g., bag, T-shirt). Most of the images contain a single person with a relatively simple pose, against relatively clean background (see Figure 5.4). **Clothing-Attribute** [144] includes 1,856 person images from social media sites, annotated with 26 attributes. Only image-level annotations are provided. However, it includes 6 texture attributes not included in Colourful-Fashion, so this method includes this auxiliary dataset mainly to enrich the representation with the 6 texture attributes.

Target Datasets: Four surveillance pedestrian datasets are used as target data. **VIPeR** [245] contains two views of 632 pedestrians. All images are normalised to 128×48 . All images are also manually labelled by [142] with 22 attributes, named VIPeR-Tag dataset [142]. **CUHK01** [246] is captured with two camera views in a campus environment. It contains 971 persons, with two images each. Images are normalised to 160×60 . **PRID450S** [247] is a recent and more realistic dataset, built on PRID 2011. It

5. Transferring a Semantic Representation

Method		VIPeR				CUHK01				PRID450S			
		r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
single	SDC [146]	25.1	44.9	56.3	70.9	15.1	25.4	31.8	40.9	23.7	38.4	46.1	58.5
	GTS [148]	25.2	50.0	62.5	75.6	-	-	-	-	-	-	-	-
	SDALF [147]	19.9	38.9	49.4	65.7	9.9	22.6	30.3	41.0	17.4	30.9	40.8	55.2
	Our unsupervised	27.7	55.3	68.3	79.7	23.3	35.8	46.6	60.7	28.5	48.9	59.6	71.3
fused	SDC_Final (eSDC) [146]	26.7	50.7	62.4	76.4	19.7	32.7	40.3	50.6	25.5	40.6	48.4	61.4
	Our unsupervised_Final	29.3	52.7	66.8	79.7	22.4	35.9	47.9	64.5	29.0	49.4	58.4	69.8

Table 5.1: Matching accuracy @ rank r (%): unsupervised learning approaches. ‘-’ indicates no result was reported and no code is available for implementation. The best results for single-cue and fused-cue methods are highlighted in bold separately.

consists of 450 image pairs recorded from two static surveillance cameras. All images are normalised to 168×80 . **PETA** [248] is a large-scale surveillance person attribute dataset that consists of 19000 images. Each image is labelled with 61 binary and 4 multi-class attributes, including colour, style etc.

Features: This method divides the image into super-pixels using a recent segmentation algorithm [105]. This method represents each super-pixel as a vector using following features: (1) Colour: 3 dimensional colour descriptors are extracted from each pixel in both RGB and LAB colour space [15, 149]. k-means is applied to obtain 150 code words for each colour space. Pixels are quantised to the nearest centres in the visual vocabulary. The resulting descriptor for each super-pixel is the normalised histogram over visual words. (2) SIFT: This method computes 128 dimensional dense SIFT over a regular grid (4×4 step size). Similar to Colour, a vocabulary of 300 words [52] is build. A histogram is built from quantised local words within each super-pixel. (3) Location: Following [52, 249], this method considers a 2 dimensional coordinate of each super-pixel centroid as an absolute location feature. A relative location feature is defined by the distances between the centroid and each of 26 human key points generated by human pose estimation [249], giving a 106 dimensional location features. The final feature vector (706D) of each super-pixel is formed by concatenation of Colour (300D), SIFT (300D) and Location (106D). To compensate for the noise in the surveillance images, this method also applies a rolling guidance filter [250] before generating super-pixels.

Settings: For training on the auxiliary datasets, the presented model uses 60 supervised factors: 34 from Colourful-Fashion (12 colour + 22 category attributes), 6 (tex-

5. Transferring a Semantic Representation

ture) from Clothing-Attribute, and 20 background factors (always off for Foreground Patch (FGP)). Thus the proposed model activates at least $K_+ \geq K_s = 60$ factors, although more may be used to explain un-annotated aspects of the data due to the use of IBP. The proposed model trains by iterating Eqs. (5.3) and (5.4) for 2000 iterations. The supervision used varies across the strongly and weakly annotated auxiliary sets. Please see appendix B for details. For transferring to the Re-ID datasets, the presented approach transfers the 60 auxiliary domain factors, and use $K_+ \geq 80$ by initialising a further 20 free factors randomly to accommodate new factors in the new domain. Any previously unseen unique aspects of the target domain can be modelled by these 20 factors. This model adapts the learned model to the target data by iterating Eqs. (5.3) and (5.5) for 100 iterations. The presented method then takes the first $K = 80$ learned factors to produce an 80-dimensional patch representation (see Section 5.3) to be used in person Re-ID.

Baselines: In addition to comparing with state-of-the-art in person Re-ID and person search methods, this chapter also considers alternative transfer methods that could potentially generate an analogous representation to the proposed framework: **SVM/MI-SVM:** SVM (as in [57, 142]) and Multi-Instance SVM [95] are used to train patch-level attribute classifiers for strongly and weakly-labelled auxiliary data respectively. The learned SVMs can then be applied to estimate feature vectors for each target image patch similarly to the proposed model. **DASA** [207]: An unsupervised domain adaptation methods to address domain shift by aligning the source and target subspaces.

Computational Cost: The complexity of the proposed algorithm is $O(JN(K^3 + KD))$ for J images with N superpixels, K factors, and D -dimensional patch features. This model run the proposed algorithm on a PC with Intel 3.47 GHz CPU and 12GB RAM. In practice this corresponds to 1 to 2 minutes for 1000 images per iteration, depending on the number of super-pixels.

5. Transferring a Semantic Representation



Figure 5.4: Visualisation of the proposed model output. Each patch is colour-coded to show the inferred dominant attribute of two types.

5.4.2 Person Re-identification

This work first evaluates person Re-ID performance against start-of-the-arts [151, 152, 155]. This experiment randomly divides the dataset into two equal, non-overlapping subsets for training and testing. This work uses the widely used accuracy at rank k of Cumulative Match Characteristic (CMC) curves to quantify performance. The results are obtained on VIPeR, CUHK01 and PRID450S datasets by averaging over 10 random splits. This work distinguishes using the suffix *_Final* the common practice of use of an ensemble of methods or features with score or feature level fusion.

Unsupervised Matching: This thesis compares the proposed model to recent state-of-the-art approaches under an unsupervised setting (i.e. no identity labels are used) including SDALF [147], eSDC [146], and GTS [148]. As shown in Table 5.1, the presented representation on its own significantly outperforms all other methods in all three datasets, and is not far off the most competitive supervised methods (Table 5.2). When fused with SDALF as in [146], performance improves further. See supplementary material for CMC curves and more comparisons.

Supervised Matching: Table 5.2 compares the proposed method in a supervised matching context against recent state-of-the-art including: MLF [161], KML [151], KISSME [152], SCNCD [149], FUSIA [165]. It shows that the proposed approach achieves comparable or better performance to state-of-the-art, especially at higher rank (i.e. $r=5,10,20$). In this setting the presented final result is obtained by fusing with kLFDA [151].

5. Transferring a Semantic Representation

Method		ViPeR				CUHK01				PRID450S			
		r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
single	MLF [161]	29.1	52.3	65.9	79.9	34.3	55.1	65.0	74.9	-	-	-	-
	KML [151]	32.3	65.8	79.7	90.9	24.0	38.9	46.7	55.4	32.4	54.4	62.4	69.6
	KISSME [152]	19.6	48.0	62.2	77.0	8.4	25.1	38.7	50.2	26.5	47.8	57.6	68.5
	SCNCD [149]	33.7	62.7	74.8	85.0	-	-	-	-	41.5	66.6	75.9	84.4
	FUSIA [165]	19.1	55.3	73.5	84.8	9.8	32.4	49.8	60.1	-	-	-	-
	Our supervised	31.1	68.6	82.8	94.9	32.7	51.2	64.4	76.3	43.1	70.5	78.2	86.3
fused	KML_Final [151]	36.1	68.7	80.1	85.6	-	-	-	-	-	-	-	-
	SCNCD_Final [149]	37.8	68.6	81.0	90.5	-	-	-	-	41.6	68.9	79.4	87.8
	MLF_Final [161]	43.4	73.0	84.9	93.7	-	-	-	-	-	-	-	-
	Our supervised_Final	41.6	71.9	86.2	95.1	31.5	52.5	65.8	77.6	44.9	71.7	77.5	86.7

Table 5.2: Matching accuracy @ rank r (%): supervised learning approaches on re-identification.

Auxiliary Data: Here this thesis evaluates the effects of various auxiliary data sources and annotations. The proposed full framework is learned with fully-supervised (f-F) Colourful-Fashion and weakly-supervised (w-C) Clothing-Attribute datasets. Table 5.3 (on ViPeR) shows that: (i) the different annotations in the two auxiliary datasets are combined synergistically, and weakly-annotated data can be used effectively (f-F+w-C > f-F); and (ii) while capable of exploiting strong supervision where available, the proposed framework does not critically rely on it (w-F+w-C close to f-F+w-C; w-F close to f-F).

Auxiliary Data	Unsupervised				Supervised			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
w-F	18.3	38.3	49.5	62.9	26.2	58.2	71.1	83.4
f-F	25.4	51.4	63.9	75.3	29.4	64.9	78.8	91.7
w-F + w-C	22.4	43.6	57.1	67.3	28.3	62.2	75.8	88.5
f-F + w-C	27.7	55.3	68.3	79.7	31.1	68.6	82.8	94.9

Table 5.3: Effects of auxiliary data source and annotation.

Contributions of Components: To evaluate the contributions of each component of the proposed framework, Table 5.4 summarises the presented model performance on ViPeR in 4 conditions: (1) Without MRF (NoMRF); (2) Direct transfer without adaptation (Eq. 5.5) (NoAdapt); (3) (1) & (2); (4) Solely unsupervised target domain learning (NoTransfer). The results show that each component (MRF modelling, transfer and

5. Transferring a Semantic Representation

adaptation) contributes to the final performance.

Method	Unsupervised				Supervised			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
NoMRF	23.3	47.4	59.1	70.9	28.0	62.1	75.2	87.2
NoAdapt	19.2	39.6	50.2	61.9	21.8	49.2	60.9	73.8
NoMrfAdapt	17.7	36.2	45.5	54.8	20.2	46.0	57.6	70.9
NoTransfer	9.5	20.5	26.9	35.6	14.3	32.9	41.7	52.5
Ours	27.7	55.3	68.3	79.7	31.1	68.6	82.8	94.9

Table 5.4: Contribution of each model component

Alternative Transfer Approaches: The proposed model is compared against alternative SVM-based approaches. Table 5.5 reveals that: (i) While (MI)SVMs can in principle deal with weakly or strongly supervised representation learning, it clearly under-performs the proposed approach, and (ii) Although conventional feature-level domain adaptation (DASA [207]) can improve the SVM performance, it is much less effective than the proposed model-level adaptation.

Method		Unsupervised				Supervised			
		r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
w-F	MI-SVM	8.0	17.8	24.4	34.4	15.6	36.2	46.5	59.9
	DASA [207]	12.2	25.8	33.9	43.7	17.1	39.2	49.4	61.5
	Ours	18.3	38.3	49.5	62.9	26.2	58.2	71.1	83.4
f-F	SVM	13.2	29.6	40.3	55.4	17.4	40.5	51.9	66.8
	DASA [207]	16.0	33.5	42.8	53.2	20.8	47.7	60.2	75.4
	Ours	25.4	51.4	63.9	75.3	29.4	64.9	78.8	91.7

Table 5.5: Comparing different transfer learning approaches

What is learned: Figure 5.4 visualises after model learning how attributes are detected given a new image. This work visualises the top 5 most confident colour and category factors/attributes for each image in the test set of Colourful-Fashion. The proposed model can almost perfectly recognise and localise the attributes (top row). As expected, the inferred attributes are much more noisy for the Re-ID data (bottom row). However,

5. Transferring a Semantic Representation

overall they are accurate (e.g. bags of different colours are detected), and crucially provide a much stronger representation than the even noisier low-level features.



Figure 5.5: Person search qualitative results. The top ranked images for each query are shown. Red boxes are false detections.

5.4.3 Person Search

Although attribute-based query is a widely studied problem, there are few studies on person search [139, 140] in surveillance. To evaluate description-based surveillance person search, this work conducts experiments on VIPeR-Tag [142] and PETA [248]. For both datasets, following [142], this experiment randomly chose 50% of the data for training (not used in the proposed transfer framework, but used in other baselines) and the remaining for testing, and repeat this procedure 10 times to obtain average results. Person Search is a retrieval task, so this work evaluates the performance of each query with a PR curve like [142, 251, 252].

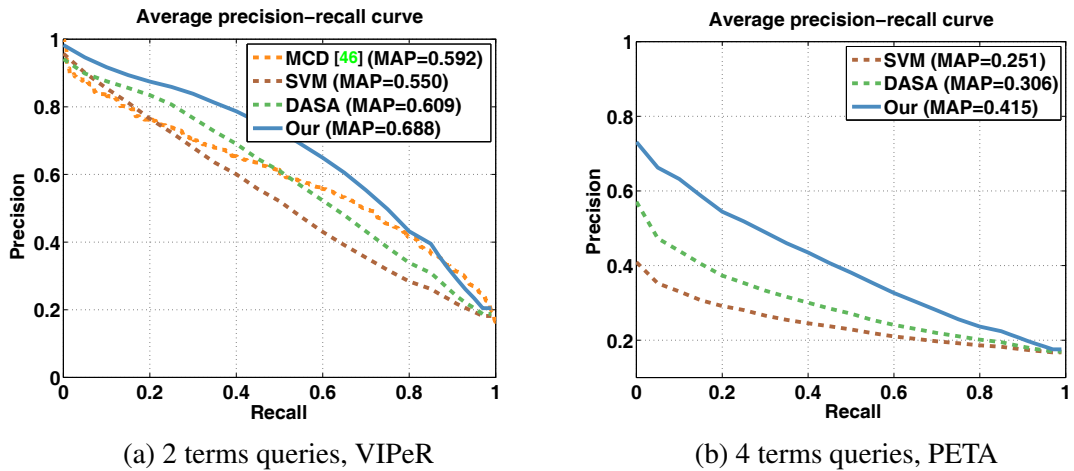


Figure 5.6: Person search: comparison with state-of-the-art.

5. Transferring a Semantic Representation

In VIPeR-Tag, all 15 queries used in [142] are contained in the source data attribute list. This thesis can thus directly compare with the results in [142]. The 15 queries are composed of a combination of an adjective (A) and a noun (N) (e.g. Red-Shirt). To ensure two query terms co-exist in the same patch, this thesis uses $\max(M_k \cdot M_{k'})$ to compute the score (see Section 5.3.2). Figure 5.6a shows the average PR curves over all annotated queries, and it is clear that the proposed method outperforms MCD [142]¹, SVM [57] and an unsupervised domain adaptation-based method DASA [207], even though no annotated VIPeR-Tag data are used for learning the proposed model. Similar to [27], SVM scores have been calibrated by [176] before being fused to generate probability estimates of queries. More detailed results (including query-specific PR curves) are available in supplementary material.

In PETA, this thesis considers a more challenging search task. Each query contains 4 terms of the form A-N+A-N (e.g. Red-Shirt+Blue-Trousers). This experiment selects all multi-class attribute labels of PETA [248], including 11 colour (A) and 4 categories (N). In total 44 A-N combinations are generated and any two of them can form a 4-term query. Like [27], the presented approach randomly generates 200 4-term queries to evaluate the methods. Note that as the query form is A-N+A-N, the two query strategies in Section 5.3.2 need to be combined to compute a score. Figure 5.6b shows that the proposed method outperforms alternatives by a larger margin in this more difficult query setting.

The proposed model has two important advantages over the compared existing methods: (1) In order to better detect conjunctive person attributes such as “Red-Shirt”, many existing methods [142] train a single attribute classifier for each combination of interest. This is not scalable because there will always be rare combinations that have too few instances to train a reliable classifier for; or at test time a combination may be required that no classifier has been trained for. By representing person attributes factorially, the proposed model has no problem searching for combinations of attributes unseen at train time. (2) Because attributes are represented conjunctively at the patch-level, this method can make complex queries such as (Black-Jeans + Blue-Shirt). An existing method such as the SVM-based one in [57], which uses image-level predictions for each attribute independently, may be confused by “Blue-Jeans + Black-Shirt”

¹Various MCD versions are evaluated in [142]. This method compares with MCD₁ which gives the best MAP.

as an alternative. This explains the larger performance margin on PETA. Figure 5.5 gives some qualitative illustration of these advantages. For example, Figure 5.5 shows that the SVM-based model in [57], learned on each attribute separately at the image level, wrongly detects a person with blue top when blue trousers is queried. This limitation is more apparent for the more challenging “Red-Shirt+Blue-Trousers” query. In contrast, with patch-based joint attribute modelling, the proposed model achieves much better results.

5.5 Summary

This chapter has introduced a framework to generate semantic attribute representations of surveillance person images for Re-ID and search. The proposed framework exploits weakly and/or strongly annotated source data from other domains and transfers it with adaptation to obtain a good representation without any target domain annotation. The resulting patch-level semantic representation obtains competitive performance for supervised Re-ID, and state-of-the-art performance for unsupervised Re-ID – which is the more practically relevant problem contexts since camera specific identity annotation is not scalable. Moreover, as a semantic representation it allows unification of Re-ID and person search within the same model.

Chapter 6

Conclusion and Future Work

This thesis has set out to explore the possibility of learning and modelling image content (including object, attributes, their locations and associations) directly from image-level weakly labelled nature images for various image understanding tasks, such as recognition, classification, image description and retrieval. In particular, the thesis is geared towards solving two problems: (1) it attempts to learn useful and reliable knowledge with minimal human annotation cost. (2) it aims to jointly model all kinds of image content (from objects to attributes) as much as possible.

To tackle this problem, Chapter 3 first presented an effective and efficient model for WSOL). The proposed model surpasses the performance of prior methods and obtains state-of-the-art results on PASCAL VOC 2007 and ImageNet datasets. It can also be applied to the YouTube-Object dataset, and to domain transfer between these image and video datasets. With joint multi-label modelling, instead of independent learning in previous work, the proposed model enables: (1) exploiting multiple object co-existence within images, (2) learning a single background shared across classes and (3) dealing with large scale data more efficiently than prior approaches. The proposed generative Bayesian formulation, enables a number of novel features: (1) integrating appearance and geometry priors, (2) exploiting inter-category appearance similarity and (3) exploiting different but related datasets via domain adaptation. Furthermore, it is able to use (potentially easier to obtain) unlabelled data with a challenging mix of relevant and irrelevant images to obtain a reasonable localiser when labelled data are in short supply for the target classes.

In order to describe objects with attributes, which is ignored in Chapter 3, an ef-

fective model for weakly-supervised learning of objects, attributes, and their locations and associations have been presented in Chapter 4. Learning object-attribute association from weak image-level labels is non-trivial but critical for learning from ‘natural’ data, and scaling to many classes and attributes. This thesis achieves this for the first time through a novel weakly-supervised stacked IBP model that simultaneously disambiguates superpixel annotation correspondence, and learns the appearance of each annotation and superpixel-level annotation correlation. The presented results show that on a variety of tasks, the proposed model often performs comparably to strongly supervised alternatives that are significantly more costly to supervise, and is consistently better than WS alternatives.

Finally, in order to utilise available cheap dataset comprehensively, chapter 5 address the WSL problem from a transfer learning perspective. It introduced a framework to generate semantic attribute representations of surveillance person images for Re-ID and search. The proposed framework exploits weakly and/or strongly annotated source data from other domains and transfers it with adaptation to obtain a good representation without any target domain annotation. The resulting patch-level semantic representation obtains competitive performance for supervised Re-ID, and state-of-the-art performance for unsupervised Re-ID – which is the more practically relevant problem contexts since camera specific identity annotation is not scalable. Moreover, as a semantic representation it allows unification of Re-ID and person search within the same model.

In summary, this thesis makes three main contributions:

1. This thesis proposes the novel concept of joint modelling of all object classes and backgrounds for WSOL. The Bayesian formulation allows it to utilise various types of prior knowledge. It also provides a solution for exploiting unlabelled data for WSOL.
2. This thesis proposes to jointly learn all object, attribute and their associations from realistic weakly labelled images. The effectiveness of the proposed WS-MRF-SIBP has been demonstrated in various image description and query tasks.
3. A transfer learning approach is proposed to learn an attribute representation from strongly or weakly annotated data or a mix. The resulting representation is useful for various surveillance applications, such as person search.

In general, this thesis showed the usefulness of top-down, cross-class and domain transfer priors – demonstrating the model’s potential to scale learning through transfer [34, 38, 91]. These contributions bring us significantly closer to the goal of scalable learning of strong models from weakly-annotated non-purpose collected data on the Internet.

Based on the work presented in this thesis, several avenues of research are possible, not only for improving the use of prior knowledge and cues of related classes, but also to tackle the problems with improved models. Possible extensions are listed below:

1. *Automatically discover* - Pre-allocated parameters or variables are always come with an assumption for expecting data distribution. These settings may be only suitable for a certain data, which is hard generalised to all types of data. Automatically determining [196] the optimal setting of parameters and variables is desired, especially for the proposed model in Chapter 3. Although the number of latent factors can be automatically discovered, there are many free parameters and variables which can be learned such as feature dimension, the number of superpixels, learning iterations, etc.
2. *Multiple layered structure* - A more hierarchical model could be developed which contain deeper multiple layer [196] to exploit parts [92, 196] and attributes [169]. These models can be applied to a specifically structured problem. For example, learning a human structure explains a human body with multiple parts, where each part show different appearance and sub-structure (head, hand). The proposed model in this thesis focuses on the daily general objects with a relatively flat model, which ignored the detailed structure of objects.
3. *Learning prior knowledge* - Prior knowledge is widely used by the proposed model in this thesis. However, most of the prior knowledge are based on the human common sense or fixed pre-defined setting such as the object-appearance similarity from pre-defined WordNet. These prior knowledge may not be suitable for different type of data. Learning prior [47] is also the interesting direction to explore.
4. *Learning from noisy data* - Although the proposed model model image content with weakly labelled natural data, it is still unclear how the proposed model deal

with noisy data. This thesis did not provide a systematic analysis on learning noisy data. In a real world, the non-purpose collected label is likely to contain realistically noisy information. It is desirable to investigate the robust capacity [169] of the proposed model or related extended models.

5. *Structure MRF* - Although the two MRFs integrated into the proposed model to capture the spatial label coherence and within-superpixel factorial correlation, another type of correlation is ignored, that is, the image-level factorial correlation which captures the global context in each image. For example, car and road typically co-exist in a street scene. Modelling such correlation provides additional constraints to learn better superpixel-level label by further disambiguating the weak image-level labels. Modelling such correlation in the current model is non-trivial and a new inference and learning algorithm needs to be developed. Another prior knowledge the current model fails to explore is the fact that each superpixel should only be explained by a single object label. Although the within-superpixel MRF can implicitly model that, the current model can be extended to suppress the co-occurrence of multiple object labels explicitly.
6. *Image language with a deep framework* - More recently, the study of automated scene understanding has moved from a single object, multiple objects, multiple objects+attributes, towards automated image captioning with full sentences [5, 26]. Most recently proposed models are hybrid of deep Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN); they are thus discriminative in nature. One of the future research directions is to integrate to the proposed model with deep learning based language models such as RNN to tackle a more challenging image captioning task whilst keeping the advantage of the introduced generative non-parametric Bayesian model for WSL.

Appendix A

Detailed Results for Object Localisation

A.1 Further Evaluations

Probabilistic feature fusion The results reported in the Chapter 3 are obtained with only SIFT features, for direct comparison with prior work. The proposed model can also exploit multiple feature fusion as additional information to further improve performance. Compared to traditional feature fusion method, the proposed probabilistic fusion (see Section 3.5 of the Chapter 3) combines multiple features at a middle level with a negligible additional computational cost. The presented method compares with two baselines:

Early fusion Arguably the simplest method is to combine features is to concatenate low-level descriptors (f) directly. The combined descriptors are then quantised into a single vocabulary. The subsequent training process is the same as using a single feature.

Late fusion This is to train completely independent models using each feature. Score level fusion (multiplication) is then applied to the different heat maps to generate the final heat map.

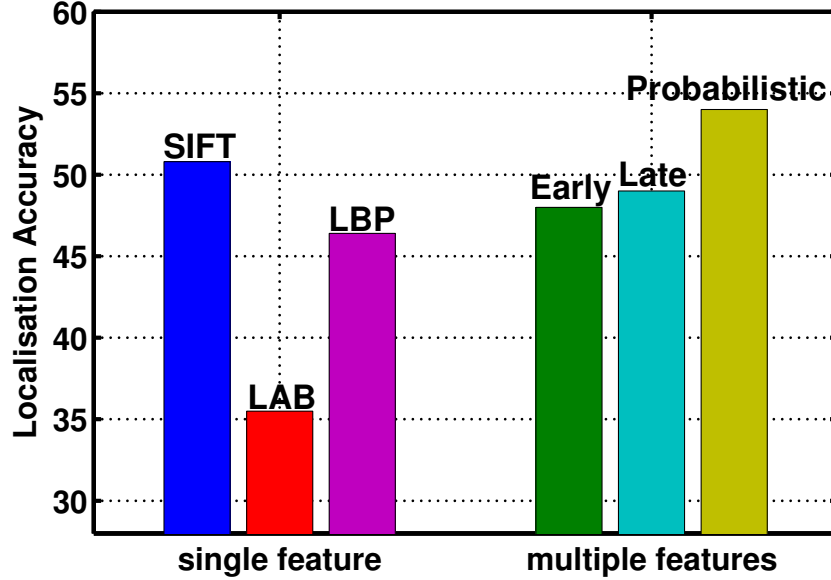


Figure A.1: Comparing different feature fusion methods.

This section evaluated the results on the VOC07-6 \times 2 dataset. In Figure A.1 it summarises the performance of multi-feature fusion and compared with the results obtained by using each single feature alone. It can be seen that colour (LAB) is a relatively weaker compared to texture based SIFT and LBP. The reason is that the colour of objects is very diverse in the challenging VOC dataset. It is noted that the simple baselines fail to combine the features constructively, with the final result worse than the best (SIFT) feature alone. In contrast, the proposed mid-level probabilistic fusion robustly and constructively combines multiple features and achieves the best performance (54%) on VOC6 \times 2.

Evaluations of individual model components Table A.1 shows how the performance of the proposed model is reduced without each of several key components, thus validating the usefulness of each of them. Specifically, IL is an analogue of the strategy of IL of each class used in existing approaches. This method trains, e.g., 12 models independently for VOC07-6 \times 2, each with only one foreground topic. These models are trained on all the data, but all instances without the target object are used as negative instances. Table A.1 verifies that this is sub-optimal compared to joint learning: each object model no longer has the benefit of utilising the other object models to explain away other foreground objects in multi-label images, thus leading to more confusion

A. Detailed Results for Object Localisation

within each image.

Without spatially aware representation (*NoSpatial*): The Gaussian representation of the location within each image enforces spatial compactness, and hence helps to disambiguate object appearance from background appearance. Without learning spatial extent, BGP of similar appearance to objects in the feature space cannot be properly disambiguated, leading to poorer learning and reduced localisation accuracy. Finally performance is also reduced without using topic-down appearance prior π^0 (*NoAppPrior*) because the model is less likely to converge to a good local minimal.

Method		VOC07-6 \times 2	VOC07-20
Our-Sampling with:	<i>IL</i>	42.5	29.8
	<i>NoSpatial</i>	44.6	32.1
	<i>NoAppPrior</i>	41.7	30.4
Alternative joint learning	<i>MIML</i> [194]	33.8	23.6
	<i>CorrLDA</i> [186]	34.3	27.2
Our-Sampling		50.8	34.1

Table A.1: Evaluation of individual components of the proposed model and comparison with alternative joint learning approaches.

Alternative joint learning approaches In this experiment, the proposed model compares other joint multi-instance/weakly-supervised multi-label learning methods, and show that none is effective for WSOL. One alternative joint learning approach is to cast WSOL as a MIML learning problem [194, 195, 253]. Most existing MIML studies consider classification. The proposed method utilises the model in [194] and reformulate it for localisation. Specifically, this method follows [42] to use the what-is-object boxes to generate bags for each image before applying MIML for localisation. Table A.1 shows that the MIML method underperforms, due to the harder discrete optimisation. This, together with being unable to integrate Bayesian prior knowledge as in the proposed model, explains its much poorer result. The presented method also compares with CorrLDA, which was designed for image annotation [186]. However its

performance is much weaker because it lacks an explicit spatial model and only admits indirect supervision of topics [185]. The proposed approach directly constrains topics via label-topic clamping, enabling more effective WS multi-label learning.

A.2 Per-Class Object Localisation Results

In Section 3.8 of the Chapter 3, it evaluated object localisation from weakly labelled data. Specifically, Table 3.1 (Chapter 3) compared the proposed methods (Our Sampling and Our Gaussian) with the state-of-the-art competitors [42, 43, 44, 91] on the three variants of the PASCAL VOC 2007 dataset: $VOC07 - 6 \times 2$, $VOC07 - 14$ and $VOC07 - 20$. Due to the space limitation, only the results averaged over all the classes in each dataset were shown. In this supplementary section, it provides more detailed per-class object localisation results for the three VOC variants in Tables A.2, A.3, and A.4 respectively. Note that few previous studies report their per-class results. Those reported per-class results are included in the tables for comparison.

As mentioned in the Chapter 3, refining the localisation by a strong detector [16, 42] brings overall improvements on the localisation accuracy. However, the improvement can be very limited or even negative for some classes when the initialisation performance is poor. For instance, Table A.4 shows that for the challenging *bottle* class, the initial WSOL accuracy is weak (6.3% for Our-Sampling+prior). After refinement using a strong detector, the localisation accuracy becomes even worse (4.2% for Our-Sampling+prior). This is understandable: with poor localisation, only a poor detector will be learned, which will not help refine the localisation.

A. Detailed Results for Object Localisation

	Initialisation			Refined by detector		
	Our-Sampling+prior	Our-Gaussian+prior	[43]	Our-Sampling+prior	Our-Gaussian+prior	[42]
aeroplane left	58.7	55.3	39.1	72.0	72.0	58
aeroplane right	64.2	72.6	50.0	71.0	71.9	59
bicycle left	29.0	34.4	28.4	60.2	58.8	46
bicycle right	36.3	38.0	30.6	48.5	48.3	40
boat left	20.7	27.8	15.1	44.4	44.2	9
boat right	27.8	19.5	20.7	46.1	48.2	16
bus left	38.1	32.8	31.0	49.7	46.0	38
bus right	52.8	47.3	35.1	61.7	54.2	74
horse left	71.4	67.1	48.5	89.8	90.1	58
horse right	69.6	77.7	45.2	85.6	88.5	52
Motorbike left	68.4	68.9	46.3	79.3	83.8	67
motorbike right	77.9	80.3	55.3	82.8	94.7	76
Average	51.2	51.8	37.1	65.9	66.7	50

Table A.2: Per-class localisation accuracy for the *VOC07-6* $\times 2$ dataset

	Initialisation		Refined by detector	
	Our-Sampling+prior	Our-Gaussian+prior	Our-Sampling+prior	Our-Gaussian+prior
aeroplane	58.8	54.1	61.0	57.2
bicycle	32.8	27.7	34.8	27.8
boat	25.8	23.4	28.5	24.9
bottle	06.8	05.7	07.3	06.4
bus	42.7	45.3	47.5	48.3
chair	06.8	06.0	09.9	07.6
diningtable	33.6	30.5	36.6	31.9
horse	57.8	48.9	58.1	54.3
motorbike	59.5	59.6	61.2	60.5
person	28.6	24.1	30.6	28.2
pottedplant	14.2	10.2	14.8	11.0
sofa	37.4	36.8	39.1	39.2
train	56.8	56.0	58.1	57.0
tvmonitor	06.5	07.4	08.4	08.3
Average	33.4	31.1	35.4	33.0

Table A.3: Per-class localisation accuracy for the *VOC07-14* dataset

A. Detailed Results for Object Localisation

	Initialisation					Refined by detector		
	Our-Sampling+prior	Our-Gaussian+prior	[43]	[90]	[91]	Our-Sampling+prior	Our-Gaussian+prior	[90]
aeroplane	62.0	53.1	38.7	45.4	54.7	68.3	63.0	42.4
bicycle	33.8	33.0	22.2	20.6	22.7	56.8	54.9	46.5
bird	32.9	20.9	27.6	29.7	33.7	37.5	24.7	18.2
boat	30.8	26.2	21.0	12.2	24.5	20.2	17.5	08.8
bottle	06.3	08.0	06.6	04.1	04.6	04.2	05.1	02.9
bus	36.5	37.8	33.3	37.1	33.9	48.8	53.7	40.9
car	42.7	41.8	39.4	41.0	42.5	63.3	42.6	73.2
cat	60.5	53.5	46.0	53.4	57.0	71.7	60.6	44.8
chair	07.4	07.6	08.1	06.5	07.3	61.0	04.6	05.4
cow	39.0	34.7	34.8	31.9	39.1	33.7	31.1	30.5
diningtable	30.4	31.7	31.5	20.5	24.1	16.2	26.4	19.0
dog	50.1	43.3	38.0	40.9	43.3	61.5	56.5	34.0
horse	57.7	51.9	37.6	37.3	41.3	55.5	54.7	48.8
motorbike	56.9	56.8	43.3	46.5	51.5	65.4	67.5	65.3
person	30.3	26.2	23.0	22.3	25.3	21.2	17.5	08.2
pottedplant	12.1	14.1	11.4	10.2	13.3	03.6	07.3	09.4
sheep	35.6	32.8	28.1	27.1	28.0	24.4	25.8	16.7
sofa	30.6	32.8	34.5	32.3	29.5	37.3	35.3	32.3
train	58.1	56.8	43.7	49.0	54.6	63.5	62.1	54.8
tvmonitor	08.2	07.0	10.5	09.8	11.8	07.8	05.7	05.5
Average	36.1	33.5	29.0	28.9	32.1	38.3	35.8	30.4

Table A.4: Per-class localisation accuracy for the *VOC07-20* dataset

A.3 Per-Class Object Detection Results

In the Chapter 3, it further evaluated object detection performance on test data given detectors trained from WS images. Table 3 (Chapter 3) reports the mAP of detection performance on both VOC-6 \times 2 and VOC-20. Here the section provides the per-class AP for VOC-6 \times 2 and VOC-20 in Table A.5 and Table A.6 respectively. It is clearly to see that for some classes (e.g. bicycle right, motorbike left in Table A.5; Bicycle, bus, car, motorbike, train in Table A.6) the proposed approach achieves comparable performance to the FS detector. This is a very encouraging result. It shows that with

A. Detailed Results for Object Localisation

our framework, automatic localisation can replace manual location annotation to train detectors for these classes. However, for those with very low localisation accuracy (e.g. bottle and pottedplant), the WS detector fails completely.

	[44]	[42]	Ours	FSL
aeroplane left	7.5	5	12.4	23
aeroplane right	21.1	18	26.2	32
bicycle left	38.5	49	48.4	59
bicycle right	44.8	62	63.6	64
boat left	0.3	0	0.2	0
boat right	0.5	0	0.5	1
bus left	0	0	0.8	21
bus right	3	16	18.2	20
horse left	45.9	29	39.3	45
horse right	17.3	14	28.5	39
Motorbike left	43.8	48	53.3	55
motorbike right	27.2	16	22.3	42
Average	20.8	21.4	26.1	33.4

Table A.5: Per-class average precision for object detection on *VOC07-6×2* dataset

A. Detailed Results for Object Localisation

	[90]	Ours	FSL
aeroplane	13.4	25.5	29.0
bicycle	44	50.0	53.6
bird	3.1	0.4	0.6
boat	3.1	9.0	13.4
bottle	0	0	26.2
bus	31.2	35.6	39.4
car	43.9	45.6	46.4
cat	7.1	14.4	16.1
chair	0.1	1.1	16.3
cow	9.3	13.4	16.5
diningtable	9.9	8.2	24.5
dog	1.5	3.2	5.0
horse	29.4	38.4	43.6
motorbike	38.3	37.3	37.8
person	4.6	16.5	35.0
pottedplant	0.1	0	8.8
sheep	0.4	2	17.3
sofa	3.8	10.0	21.6
train	34.2	34.0	34.0
tvmonitor	0	0.2	39.0
Average	13.9	17.2	26.3

Table A.6: Per-class average precision for object detection on *VOC07-20* dataset

Appendix B

Detailed Results for Re-identification and Search

B.1 Details on Supervision

In Section 5.4 of the thesis, under the heading “Settings”, it has briefly mentioned that the supervision used for the proposed model varies across the strongly and weakly annotated auxiliary sets. This section now gives the details on the different types of annotations and corresponding supervision used in the presented experiments in Table B.1.

	Full / Fashion [52]			Weak / Clothing [144]	
Factor	Color	Category	BG	Texture	BG
FGP	Strong	Strong	$L_k = 0$	Weak	None
BGP	None	$L_k = 0$	None		

Table B.1: Different types of supervision used by the proposed models depend on how different attributes of each auxiliary dataset are annotated.

In the auxiliary datasets, there are various types of attributes (modelled as factors in the proposed model) and each can be annotated in different ways (denoted as L_k , see Section 5.4 of the thesis). More specifically, there are three factor types: Colour (12

B. Detailed Results for Re-identification and Search

factors), Category (22), Texture (6) and Background (BG) (20). The number of factors depend on the actual annotation from [52, 144]. These factors may be supervised strongly (L_k given by pixel level annotation), weakly (L_k image level annotation) or not at all (All $L_k = 1$).

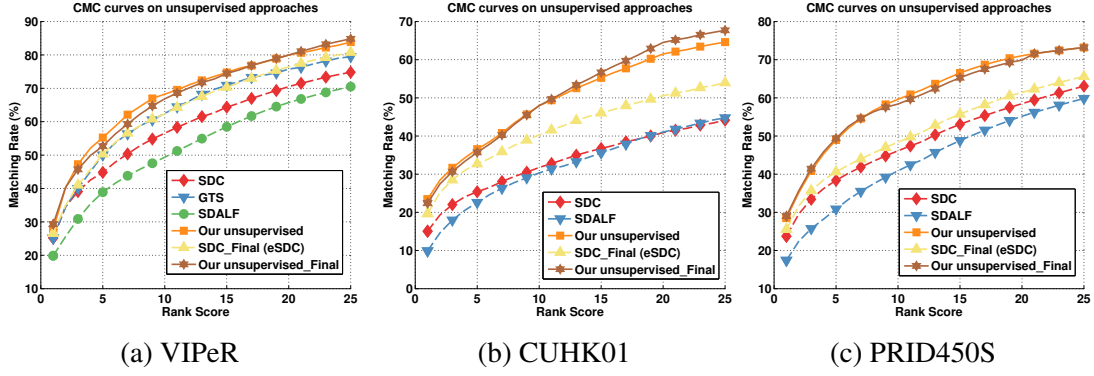


Figure B.1: CMC comparison of unsupervised learning based approaches.

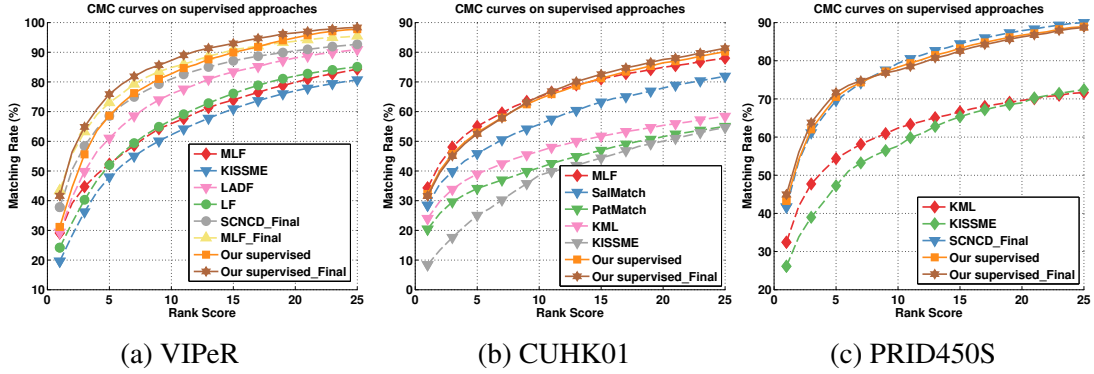


Figure B.2: CMC comparison of supervised learning based approaches.

Given strongly supervised data (Colourful-Fashion [52]), all superpixels/patches can be categorised into FGP and BGP (Table B.1 rows). FGP cannot contain background factors ($L_k = 0$). Colour and category data are strongly annotated so the learning of these factors are strongly supervised for the foreground (Strong). Colour factors can also occur on the background, so these are not supervised (None). Meanwhile, Category factors (e.g., shirt) cannot occur on the background, so $L_k = 0$ here. Finally, any background factors are free to be used on any BGP (None).

Given a WS dataset (Clothing-Attribute [144]), the proposed method does not know the location of foreground background pixels, so there is no patch-type break-

down. Therefore the texture factors are WS (Weak); meanwhile the background factors are not supervised.

B.2 Re-identification Performance Measured by CMC

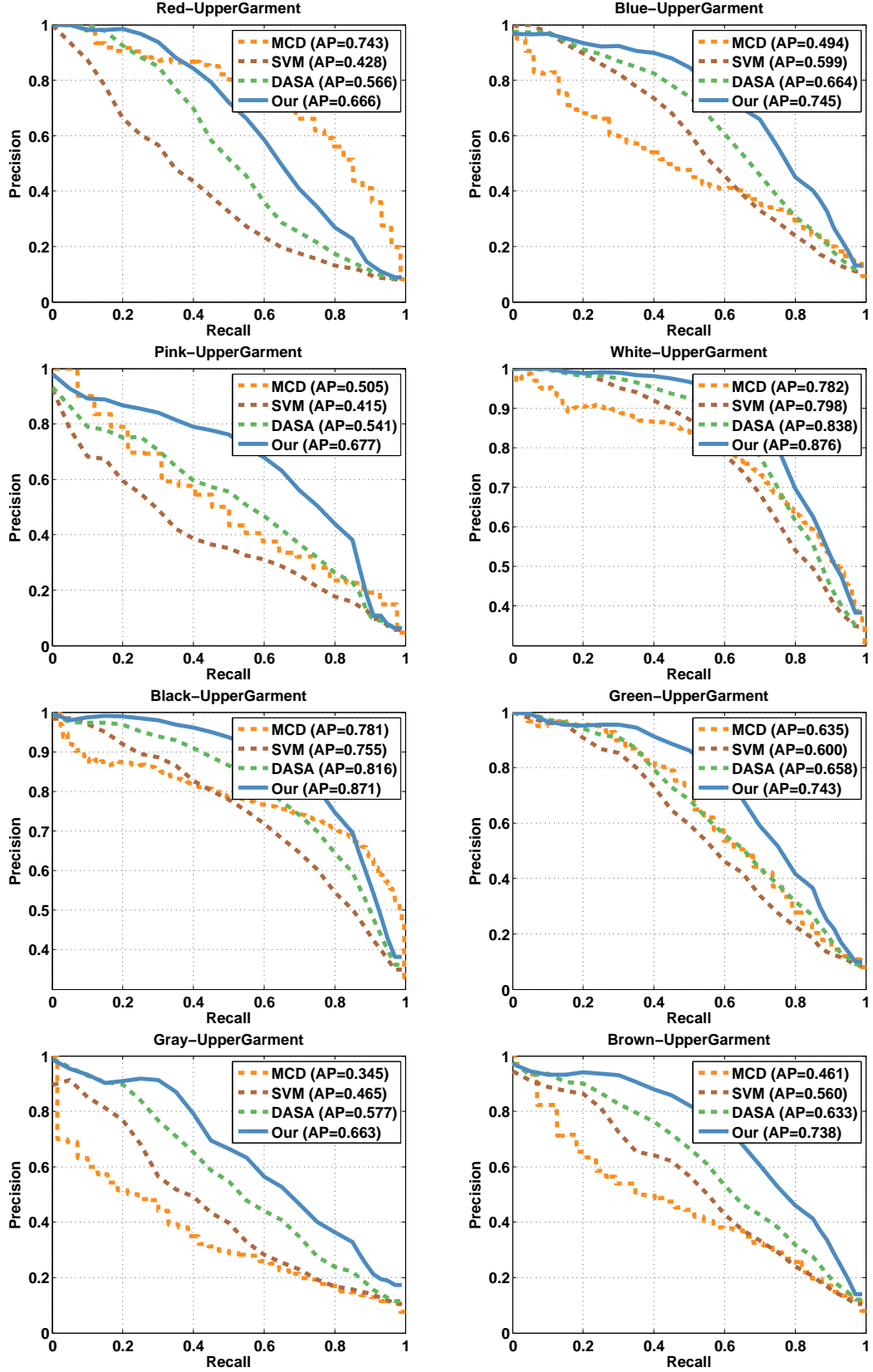
In Table 5.1 and Table 5.2 (Section 5.4) of the thesis, the proposed methods are compared with existing unsupervised and supervised learning approaches respectively. Due to space limitation, only the matching accuracy @ rank [1, 5, 10, 20] are reported. In this supplementary material, it provides the CMC curves over rank 1 to 25 obtained by different compared methods on the three datasets. Figure B.1 shows the CMC curves of the compared unsupervised learning based methods, including SDC [146], GTS [148], SDALF [147] and SDC_Final (eSDC) [146]. Figure B.2 compares the supervised learning based methods, including MLF [161], KISSME [152], LADF [160], LF [156], SCNCD_Final [149], PatMatch [153], SalMatch [153], KML [151] and MLF_Final [161]. Note that the CMC curves of some compared existing methods (i.e. KISSME [152], KML [151], SDC [146] and SDALF [147]) on some datasets are obtained based on our own implementation when their CMC curves were not reported in previous works. In addition, the CMC curves for some methods are not included in these plots if previous work provides neither the CMC curves nor the code for us to implement.

B.3 Per-query Person Search Results

In Figure 5.6 of the thesis, it reported the average PR curves of all queries for various person search methods. Here it provides more details on their person search performance on the VIPeR-Tag [142] dataset in the form of the precision-recall (PR) curve of each query. There are 200 random queries used on the PETA [248] dataset; it is thus not possible to present the per-query PR curves on PETA here.

These per-query PR curves on VIPeR-Tag are shown in Figure B.3. Note that the queries used on this dataset as defined in [142] have different attribute names as those in the auxiliary Colourful-Fashion dataset. For examples, the attributes ‘UpperGarment’

B. Detailed Results for Re-identification and Search



B. Detailed Results for Re-identification and Search

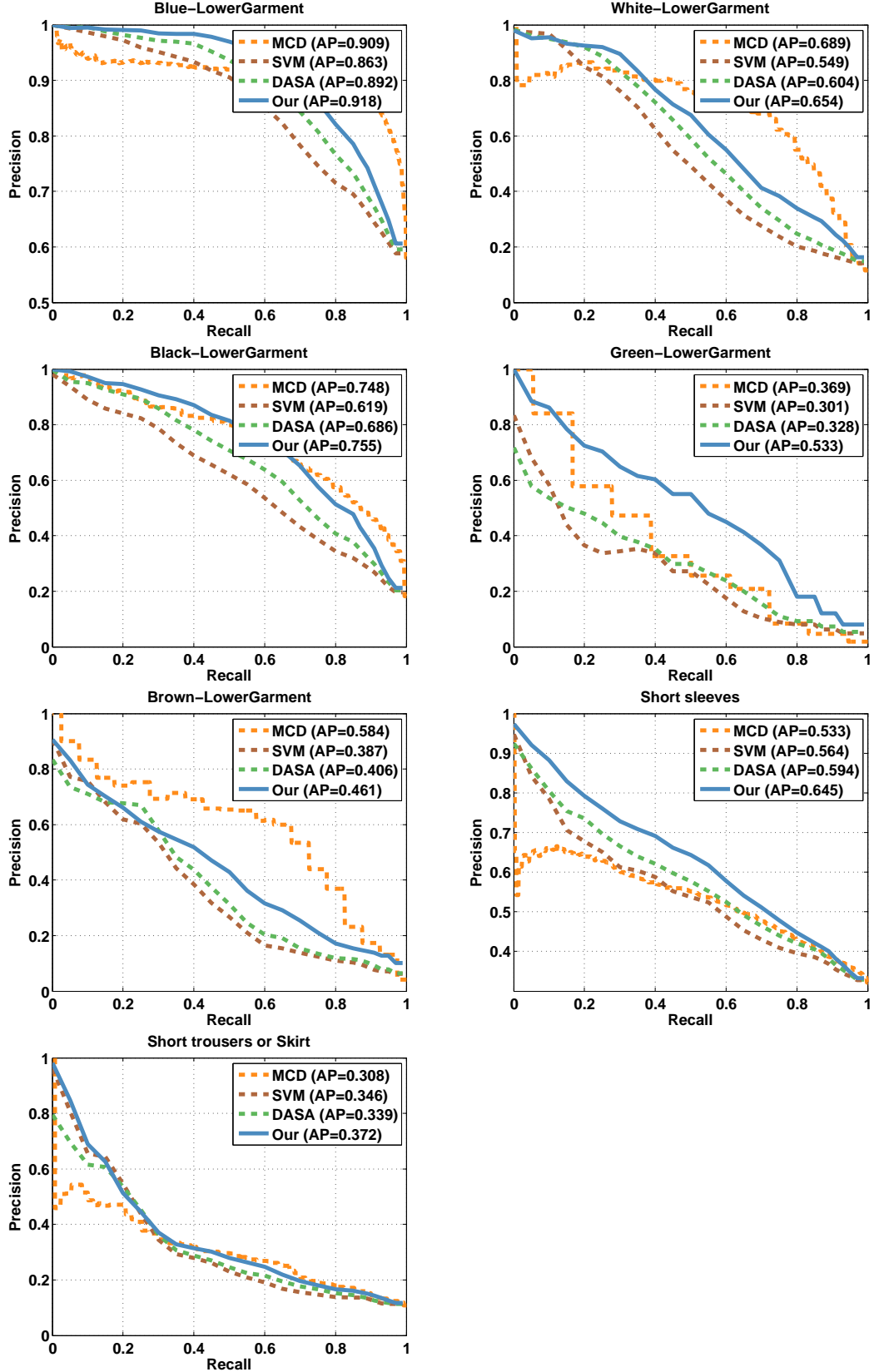


Figure B.3: Person search performance on each object-attribute query

B. Detailed Results for Re-identification and Search

and ‘LowerGarment’ appear in 13 out of the 15 queries. However, much finer-grained attributes are used in the Colourful-Fashion dataset; for instance, for ‘UpperGarment’, it has ‘blazer’, ‘T-shirt’, ‘blouse’, and ‘sweater’. These finer-grained attributes are thus merged to form the two coarser-grained attributes on VIPeR-Tag before person search.

References

- [1] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg, “Baby talk: Understanding and generating simple image descriptions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [x](#), [1](#), [2](#), [33](#), [34](#), [85](#), [93](#)
- [2] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, “segdeepm: Exploiting segmentation and context in deep neural networks for object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [x](#), [3](#), [4](#)
- [3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [x](#), [4](#), [5](#)
- [4] V. Ferrari and A. Zisserman, “Learning visual attributes,” in *Advances in Neural Information Processing Systems*, 2008. [x](#), [4](#), [6](#), [32](#)
- [5] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [x](#), [7](#), [34](#), [120](#)
- [6] B. Siddiquie, R. Feris, and L. Davis, “Image ranking and retrieval based on multi-attribute queries,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [x](#), [7](#), [8](#), [33](#)
- [7] I. Biederman, “Human image understanding: Recent research and a theory,” *Computer Vision, Graphics, and Image Processing*, vol. 32, no. 1, pp. 29 – 73, 1985. [1](#)

REFERENCES

- [8] D. Lowe, “Object recognition from local scale-invariant features,” in *IEEE International Conference on Computer Vision*, vol. 2, no. 1, 1999, pp. 1150–1157. [1](#)
- [9] P. Allen, A. Timcenko, B. Yoshimi, and P. Michelman, “Automated tracking and grasping of a moving object with a robotic hand-eye system,” *IEEE Transactions on Robotics and Automation*, vol. 9, no. 2, pp. 152–165, 1993. [1](#), [6](#)
- [10] A. Stent and A. Loui, “Using event segmentation to improve indexing of consumer photographs,” in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, vol. 1, no. 7, 2001, pp. 59–65. [1](#), [6](#)
- [11] T. Xiang and S. Gong, “Beyond tracking: Modelling activity and understanding behaviour,” *International Journal of Computer Vision*, vol. 67, no. 1, pp. 21–51, 2006. [1](#), [6](#)
- [12] C. Pope and K. Kaur, “Is it human or computer? defending e-commerce with captchas,” *IT Professional*, vol. 7, no. 2, pp. 43–49, 2005. [1](#), [6](#)
- [13] G. Heitz and D. Koller, “Learning spatial context: Using stuff to find things,” in *European Conference on Computer Vision*, 2008. [2](#)
- [14] M. Pontil and A. Verri, “Support vector machines for 3d object recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637–646, 1998. [2](#)
- [15] T. Gevers and A. Smeulders, “Color based object recognition,” *Pattern Recognition*, 1997. [2](#), [109](#)
- [16] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. [3](#), [10](#), [25](#), [42](#), [50](#), [58](#), [61](#), [62](#), [85](#), [124](#)
- [17] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001. [3](#)

REFERENCES

- [18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results,” <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 3, 10, 54
- [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 3, 54
- [20] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000. 4
- [21] J. Carreira and C. Sminchisescu, “Cpmc: Automatic object segmentation using constrained parametric min-cuts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1312–1328, 2012. 4, 27, 28
- [22] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 4, 30, 33, 34, 84, 85, 87, 90, 91, 92
- [23] D. Hanwell and M. Mirmehdi, “Weakly supervised learning of semantic colour terms,” *IET Computer Vision*, vol. 8, no. 2, pp. 110–117, 2014. 4, 29
- [24] G. Anitha and R. G. Kumar, “Vision based autonomous landing of an unmanned aerial vehicle,” *Procedia Engineering*, vol. 38, no. 1, pp. 2250 – 2256, 2012. 6
- [25] G. Yahav, G. Iddan, and D. Mandelboum, “3d imaging camera for gaming application,” in *International Conference on Consumer Electronics*, 2007. 6
- [26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 7, 34, 120
- [27] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi, “Multi-attribute queries: To merge or not to merge?” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 7, 33, 90, 115

REFERENCES

- [28] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, “What does classifying more than 10,000 image categories tell us?” in *European Conference on Computer Vision*, 2010. [9](#), [42](#)
- [29] T. Hospedales, J. Li, S. Gong, and T. Xiang, “Identifying rare and subtle behaviors: A weakly supervised joint topic model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2451–2464, 2011. [9](#), [35](#), [42](#)
- [30] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, “Image segmentation with a bounding box,” in *IEEE International Conference on Computer Vision*, 2009. [9](#), [10](#), [42](#)
- [31] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, “The re-identification challenge,” in *Person Re-Identification*. Springer London, 2014. [9](#), [19](#)
- [32] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother, “Weakly supervised discriminative localization and classification: a joint learning process,” in *IEEE International Conference on Computer Vision*, 2009. [10](#), [15](#), [26](#), [27](#), [42](#)
- [33] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu, “Multiple Component Learning for Object Detection,” in *European Conference on Computer Vision*, 2008. [10](#)
- [34] D. Kuettel, M. Guillaumin, and V. Ferrari, “Segmentation propagation in imagenet,” in *European Conference on Computer Vision*, 2012. [10](#), [37](#), [42](#), [71](#), [119](#)
- [35] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, “Unsupervised joint object discovery and segmentation in internet images,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [10](#), [92](#)
- [36] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013. [10](#), [84](#)

REFERENCES

- [37] W. Ouyang, H. Li, X. Zeng, and X. Wang, “Learning deep representation with large-scale attributes,” in *IEEE International Conference on Computer Vision*, 2015. [10](#)
- [38] M. Guillaumin and V. Ferrari, “Large-scale Knowledge Transfer for Object Localization in ImageNet,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [11](#), [15](#), [26](#), [37](#), [38](#), [42](#), [63](#), [71](#), [119](#)
- [39] J. Yang, R. Yan, and A. G. Hauptmann, “Cross-domain video concept detection using adaptive svms,” in *International Conference on Multimedia*, 2007. [14](#), [37](#), [53](#)
- [40] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, “Learning object class detectors from weakly annotated video,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [14](#), [26](#), [38](#), [54](#), [55](#), [64](#), [65](#), [66](#)
- [41] A. Torralba and A. Efros, “Unbiased look at dataset bias,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [14](#), [53](#)
- [42] T. Deselaers, B. Alexe, and V. Ferrari, “Weakly supervised localization and learning with generic knowledge,” *International Journal of Computer Vision*, vol. 100, no. 3, pp. 275–293, 2012. [15](#), [26](#), [34](#), [36](#), [37](#), [38](#), [39](#), [52](#), [54](#), [56](#), [57](#), [62](#), [70](#), [123](#), [124](#), [125](#), [127](#)
- [43] P. Siva, C. Russell, and T. Xiang, “In defence of negative mining for annotating weakly labelled data,” in *European Conference on Computer Vision*, 2012. [15](#), [26](#), [36](#), [52](#), [54](#), [56](#), [57](#), [70](#), [124](#), [125](#), [126](#)
- [44] M. Pandey and S. Lazebnik, “Scene recognition and weakly supervised object localization with deformable part-based models,” in *IEEE International Conference on Computer Vision*, 2011. [15](#), [26](#), [27](#), [36](#), [38](#), [52](#), [54](#), [56](#), [57](#), [58](#), [62](#), [124](#), [127](#)
- [45] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, “Weakly supervised object recognition and localization with stable segmentations,” in *European Conference on Computer Vision*, 2008. [15](#)

REFERENCES

- [46] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, “Co-localization in real-world images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [15](#), [26](#), [54](#), [55](#), [56](#), [57](#), [63](#)
- [47] R. Salakhutdinov, A. Torralba, and J. Tenenbaum, “Learning to share visual appearance for multiclass object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [15](#), [37](#), [119](#)
- [48] X. Zhu, “Semi-supervised learning literature survey.” University of Wisconsin-Madison Department of Computer Science, Tech. Rep. 1530, 2007. [16](#), [38](#)
- [49] T. L. Griffiths and Z. Ghahramani, “The indian buffet process: An introduction and review,” *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011. [18](#), [21](#), [35](#), [102](#), [105](#), [106](#)
- [50] S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds., *Person Re-Identification*. Springer, 2014. [19](#), [31](#)
- [51] R. Vezzani, D. Baltieri, and R. Cucchiara, “People re-identification in surveillance and forensics: a survey,” *ACM Computing Surveys*, vol. 46, no. 2, pp. 29:1–29:37, 2013. [19](#), [31](#)
- [52] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, “Fashion parsing with weak color-category labels,” *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 253–265, 2014. [20](#), [31](#), [32](#), [108](#), [109](#), [129](#), [130](#)
- [53] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, “Parsing clothing in fashion photographs.” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [20](#), [32](#)
- [54] W. Yang, P. Luo, and L. Lin, “Clothing co-parsing by joint image segmentation and labeling,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [20](#)
- [55] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg, “Paper doll parsing: Retrieving similar styles to parse clothing items,” in *IEEE International Conference on Computer Vision*, 2013. [20](#), [31](#)

REFERENCES

- [56] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, “Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [20](#)
- [57] R. Layne, T. Hospedales, and S. Gong, *Person Re-identification*. Springer, 2014, ch. Attributes-based Re-identification. [21](#), [31](#), [32](#), [102](#), [110](#), [115](#), [116](#)
- [58] J. Roth and X. Liu, “On the exploration of joint attribute learning for person re-identification,” in *Asian Conference on Computer Vision*, 2014. [21](#), [32](#), [102](#)
- [59] J. Feng, S. Jegelka, S. Yan, and T. Darrell, “Learning scalable discriminative dictionary with sample relatedness,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [21](#), [33](#), [103](#)
- [60] Z. Shi, Y. Yang, T. M. Hospedales, and T. Xiang, “Weakly supervised learning of objects, attributes and their associations,” in *European Conference on Computer Vision*, 2014. [21](#), [103](#)
- [61] K. Palla, D. Knowles, and Z. Ghahramani, “Relational learning and network modelling using infinite latent attribute models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 462–474, 2014. [21](#), [103](#)
- [62] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. [24](#), [25](#)
- [63] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, “Pedestrian detection using wavelet templates,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 1997. [24](#)
- [64] W. Nam, P. Dollar, and J. H. Han, “Local decorrelation for improved pedestrian detection,” in *Advances in Neural Information Processing Systems*, 2014. [24](#)
- [65] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in *European Conference on Computer Vision*, 2014. [24](#)
- [66] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li, “Object detection by labeling superpixels,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [24](#)

REFERENCES

- [67] C. Gu, J. Lim, P. Arbelaez, and J. Malik, “Recognition using regions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [24](#)
- [68] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, “Selective search for object recognition,” *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013. [24](#)
- [69] B. Alexe, T. Deselaers, and V. Ferrari, “Measuring the objectness of image windows,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012. [24](#), [26](#), [56](#), [63](#)
- [70] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004. [24](#)
- [71] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [24](#), [25](#)
- [72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [24](#)
- [73] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *European Conference on Computer Vision*, 2014. [24](#)
- [74] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, “Bing: Binarized normed gradients for objectness estimation at 300fps,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [24](#)
- [75] S. Manen, M. Guillaumin, and L. Van Gool, “Prime object proposals with randomized prim’s algorithm,” in *IEEE International Conference on Computer Vision*, 2013. [24](#)
- [76] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, “Multiscale combinatorial grouping,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [24](#)

REFERENCES

- [77] J. Hosang, R. Benenson, and B. Schiele, “How good are detection proposals, really?” in *British Machine Vision Conference*, 2014. [24](#)
- [78] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. [25](#)
- [79] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering object categories in image collections,” in *IEEE International Conference on Computer Vision*, 2005. [25](#), [35](#), [43](#)
- [80] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods,” in *British Machine Vision Conference*, 2011. [25](#)
- [81] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *European Conference on Computer Vision*, 2010. [25](#)
- [82] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting,” *Annals of Statistics*, vol. 28, no. 2, pp. 337–374, 1998. [25](#)
- [83] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. [25](#)
- [84] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, “Support vector machine learning for interdependent and structured output spaces,” in *International Conference on Machine Learning*, 2004. [25](#)
- [85] T. Joachims, T. Finley, and C.-N. Yu, “Cutting-plane training of structural svms,” *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009. [25](#)
- [86] C.-N. J. Yu and T. Joachims, “Learning structural svms with latent variables,” in *International Conference on Machine Learning*, 2009. [25](#)
- [87] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2012. [25](#)

REFERENCES

- [88] R. Girshick, F. Iandola, T. Darrell, and J. Malik, “Deformable part models are convolutional neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [25](#)
- [89] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 1, no. 1, pp. 1–42, 2015. [25](#)
- [90] P. Siva and T. Xiang., “Weakly supervised object detector learning with model drift detection,” in *IEEE International Conference on Computer Vision*, 2011. [26](#), [27](#), [39](#), [52](#), [54](#), [56](#), [57](#), [58](#), [62](#), [70](#), [126](#), [128](#)
- [91] Z. Shi, P. Siva, and T. Xiang, “Transfer learning by ranking for weakly supervised object annotation,” in *British Machine Vision Conference*, 2012. [26](#), [52](#), [54](#), [56](#), [57](#), [71](#), [119](#), [124](#), [126](#)
- [92] D. Crandall and D. Huttenlocher, “Weakly supervised learning of part-based spatial models for visual object recognition,” in *European Conference on Computer Vision*, 2006. [26](#), [119](#)
- [93] A. Joulin, K. Tang, and L. Fei-Fei, “Efficient image and video co-localization with frank-wolfe algorithm,” in *European Conference on Computer Vision*, 2014. [26](#), [64](#), [65](#)
- [94] O. Maron and T. Lozano-Prez, “A framework for multiple-instance learning,” in *Advances in Neural Information Processing Systems*, 1998. [26](#)
- [95] S. Andrews, I. Tsochantaridis, and T. Hofmann, “Support vector machines for multiple-instance learning,” in *Advances in Neural Information Processing Systems*, 2003. [26](#), [110](#)
- [96] N. Nguyen, “A new svm approach to multi-instance multi-label learning,” in *IEEE International Conference on Data Mining*, 2010. [26](#), [34](#), [36](#)
- [97] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, “Shifting weights: Adapting object detectors from image to video,” in *Advances in Neural Information Processing Systems*, 2012. [26](#), [38](#), [53](#), [66](#)

REFERENCES

- [98] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object?” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [27](#), [36](#)
- [99] O. Chum and A. Zisserman, “An exemplar model for learning object classes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. [27](#)
- [100] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman, “Using multiple segmentations to discover objects and their extent in image collections,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. [27](#)
- [101] T. Deselaers, B. Alexe, and V. Ferrari, “Localizing objects while learning their appearance,” in *European Conference on Computer Vision*, 2010. [27](#)
- [102] J. Muerle and D. Alle, “Experimental evaluation of techniques for automatic segmentation of objects in a complex scene,” *Pictorial Pattern Recognition*, vol. 5, no. 2, pp. 323–339, 1968. [27](#)
- [103] M. Mirmehdi and M. Petrou, “Segmentation of color textures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, pp. 142–159, 2000. [27](#)
- [104] J. Kim and K. Grauman, “Shape sharing for object segmentation,” in *European Conference on Computer Vision*, 2012. [27](#), [28](#)
- [105] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011. [27](#), [74](#), [109](#)
- [106] D. Weiss and B. Taskar, “Scalpel: Segmentation cascades with localized priors and efficient learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [27](#), [28](#)
- [107] J. Carreira, F. Li, and C. Sminchisescu, “Object recognition by sequential figure-ground ranking,” *International Journal of Computer Vision*, vol. 98, no. 3, pp. 243–262, 2012. [28](#)

REFERENCES

- [108] A. Sharma, O. Tuzel, and D. W. Jacobs, “Deep hierarchical parsing for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 28
- [109] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” in *IEEE International Conference on Computer Vision*, 2009. 28
- [110] D. Munoz, J. A. Bagnell, and M. Hebert, “Stacked hierarchical labeling,” in *European Conference on Computer Vision*, 2010. 28
- [111] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh, “Analyzing semantic segmentation using hybrid human-machine crfs,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 28
- [112] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “The role of context for object detection and semantic segmentation in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 28
- [113] V. Lempitsky, A. Vedaldi, and A. Zisserman, “Pylon model for semantic segmentation,” in *Advances in Neural Information Processing Systems*, 2011. 28
- [114] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013. 28
- [115] P. H. O. Pinheiro and R. Collobert, “Recurrent convolutional neural networks for scene labeling,” in *International Conference on Machine Learning*, 2014. 28
- [116] J. Tighe and S. Lazebnik, “Superparsing: Scalable nonparametric image parsing with superpixels,” in *European Conference on Computer Vision*, 2010. 29
- [117] C. Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing via label transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368–2382, 2011. 29, 91, 92

REFERENCES

- [118] G. Singh and J. Kosecka, “Nonparametric scene parsing with adaptive feature relevance and semantic context,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [29](#), [95](#), [96](#)
- [119] D. Eigen and R. Fergus, “Nonparametric image parsing using adaptive neighbor sets,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [29](#)
- [120] J. Tighe and S. Lazebnik, “Finding things: Image parsing with regions and per-exemplar detectors,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [29](#)
- [121] J. Yang, B. Price, S. Cohen, and M.-H. Yang, “Context driven scene parsing with attention to rare classes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [29](#), [95](#), [96](#)
- [122] J. Tighe, M. Niethammer, and S. Lazebnik, “Scene parsing with object instance inference using regions and per-exemplar detectors,” *International Journal of Computer Vision*, vol. 112, no. 2, pp. 150–171, 2015. [29](#), [95](#), [96](#)
- [123] S. Zheng, M.-M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C. Rother, and P. Torr, “Dense semantic image segmentation with objects and attributes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [29](#), [30](#), [91](#), [92](#), [93](#)
- [124] J. Xu, A. G. Schwing, and R. Urtasun, “Tell me what you see and i will show you where it is,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [29](#), [34](#), [83](#), [95](#), [97](#)
- [125] A. Vezhnevets, V. Ferrari, and J. Buhmann, “Weakly supervised structured output learning for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [29](#), [34](#), [95](#)
- [126] M. Rubinstein, C. Liu, and W. T. Freeman, “Annotation propagation in large image databases via dense image correspondence,” in *European Conference on Computer Vision*, 2012. [29](#), [34](#)

REFERENCES

- [127] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, “Weakly-supervised dual clustering for image semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [29](#), [34](#), [83](#), [86](#), [87](#), [92](#), [93](#), [95](#)
- [128] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, “Unsupervised joint object discovery and segmentation in internet images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [29](#)
- [129] Z. Li, E. Gavves, T. Mensink, and C. G. Snoek, “Attributes make sense on segmented objects,” in *European Conference on Computer Vision*, 2014. [30](#), [93](#)
- [130] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by betweenclass attribute transfer,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [30](#)
- [131] D. Mahajan, S. Sellamanickam, and V. Nair, “A joint learning framework for attribute models and object descriptions,” in *IEEE International Conference on Computer Vision*, 2011. [30](#), [33](#)
- [132] Y. Wang and G. Mori, “A discriminative latent model of object classes and attributes,” in *European Conference on Computer Vision*, 2010. [30](#), [33](#)
- [133] D. Jayaraman, F. Sha, and K. Grauman, “Decorrelating semantic visual attributes by resisting the urge to share,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [30](#)
- [134] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for attribute-based classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [30](#)
- [135] F. Song, X. Tan, and S. Chen, “Exploiting relationship between attributes for improved face verification,” in *British Machine Vision Conference*, 2012. [30](#)
- [136] D. Parikh and K. Grauman, “Relative attributes,” in *IEEE International Conference on Computer Vision*, 2011. [30](#)

REFERENCES

- [137] K. Duan, D. Parikh, D. Crandall, and K. Grauman, “Discovering localized attributes for fine-grained recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [30](#), [31](#)
- [138] S. Wang, J. Joo, Y. Wang, and S.-C. Zhu, “Weakly supervised learning for attribute localization in outdoor scenes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [30](#), [33](#), [34](#), [85](#)
- [139] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, “Attribute-based people search in surveillance environments,” in *WACV*, 2009. [31](#), [32](#), [114](#)
- [140] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, “Attribute-based people search: Lessons learnt from a practical surveillance system,” in *International Conference on Multimedia Retrieval*, 2014. [31](#), [114](#)
- [141] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, “Pose search: Retrieving people using their pose,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [31](#)
- [142] R. Satta, G. Fumera, and F. Roli, “People search with textual queries about clothing appearance attributes,” in *Person Re-Identification*. Springer, 2014. [31](#), [32](#), [108](#), [110](#), [114](#), [115](#), [131](#)
- [143] N. Kumar, P. Belhumeur, and S. Nayar, “Facetracer: A search engine for large collections of images with faces,” in *European Conference on Computer Vision*, 2008. [31](#), [32](#)
- [144] H. Chen, A. Gallagher, and B. Girod, “Describing clothing by semantic attributes,” in *European Conference on Computer Vision*, 2012. [31](#), [32](#), [33](#), [108](#), [129](#), [130](#)
- [145] L. Bourdev, S. Maji, and J. Malik, “Describing people: Poselet-based attribute classification,” in *IEEE International Conference on Computer Vision*, 2011. [31](#)
- [146] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [31](#), [109](#), [111](#), [131](#)

REFERENCES

- [147] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [31](#), [109](#), [111](#), [131](#)
- [148] H. Wang, S. Gong, and T. Xiang, “Unsupervised learning of generative topic saliency for person re-identification,” in *British Machine Vision Conference*, 2014. [31](#), [33](#), [109](#), [111](#), [131](#)
- [149] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Li, “Salient color names for person re-identification,” in *European Conference on Computer Vision*, 2014. [31](#), [109](#), [111](#), [112](#), [131](#)
- [150] A. Das, A. Chakraborty, and A. Roy-Chowdhury, “Consistent re-identification in a camera network,” in *European Conference on Computer Vision*, 2014. [31](#)
- [151] F. Xiong, M. Gou, O. Camps, and M. Sznai, “Person re-identification using kernel-based metric learning methods,” in *European Conference on Computer Vision*, 2014. [31](#), [32](#), [107](#), [108](#), [111](#), [112](#), [131](#)
- [152] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [31](#), [32](#), [111](#), [112](#), [131](#)
- [153] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by salience matching,” in *IEEE International Conference on Computer Vision*, 2013. [31](#), [32](#), [131](#)
- [154] C.-H. Kuo, S. Khamis, and V. Shet, “Person re-identification using semantic color names and rankboost,” in *IEEE Workshop on Applications of Computer Vision*, 2013. [31](#)
- [155] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *European Conference on Computer Vision*, 2008. [31](#), [111](#)
- [156] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, “Local fisher discriminant analysis for pedestrian re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [31](#), [131](#)

REFERENCES

- [157] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [31](#)
- [158] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, “Semi-supervised coupled dictionary learning for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [32](#)
- [159] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, “Human re-identification by matching compositional template with cluster sampling,” in *IEEE International Conference on Computer Vision*, 2013. [32](#)
- [160] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, “Learning locally-adaptive decision functions for person verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [32](#), [131](#)
- [161] R. Zhao, W. Ouyang, and X. Wang, “Learning mid-level filters for person re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [32](#), [111](#), [112](#), [131](#)
- [162] A. J. Ma, P. C. Yuen, and J. Li, “Domain transfer support vector ranking for person re-identification without target camera label information,” in *IEEE International Conference on Computer Vision*, 2013. [32](#)
- [163] W.-S. Zheng, S. Gong, and T. Xiang, “Transfer re-identification: From person to set-based verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [32](#)
- [164] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, “Attribute-restricted latent topic model for person re-identification,” *Pattern Recognition*, vol. 45, no. 12, pp. 4204–4213, 2012. [32](#)
- [165] R. Layne, T. Hospedales, and S. Gong, “Re-id: Hunting attributes in the wild,” in *British Machine Vision Conference*, 2014. [32](#), [111](#), [112](#)
- [166] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, “Clothing attributes assisted person re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869–878, 2014. [32](#)

REFERENCES

- [167] S. Khamis, C.-H. Kuo, V. K. Singh, V. Shet, and L. S. Davis, “Joint learning for attribute-consistent person re-identification,” in *European Conference on Computer Vision Workshop on Visual Surveillance and Re-Identification*, 2014. [32](#)
- [168] C. L. Zitnick and D. Parikh, “Bringing semantics into focus using visual abstraction,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [32](#)
- [169] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, “Attribute learning for understanding unstructured social activity,” in *European Conference on Computer Vision*, 2012. [32](#), [33](#), [35](#), [119](#), [120](#)
- [170] N. Turakhia and D. Parikh, “Attribute dominance: What pops out?” in *IEEE International Conference on Computer Vision*, 2013. [33](#)
- [171] L. Bourdev, S. Maji, and J. Malik, “Describing people: A poselet-based approach to attribute classification,” in *IEEE International Conference on Computer Vision*, 2011. [33](#), [34](#)
- [172] G. Wang and D. Forsyth, “Joint learning of visual attributes, object classes and visual saliency,” in *IEEE International Conference on Computer Vision*, 2009. [33](#), [85](#)
- [173] A. Kovashka, S. Vijayanarasimhan, and K. Grauman, “Actively selecting annotations among objects and attributes,” in *IEEE International Conference on Computer Vision*, 2011. [33](#)
- [174] X. Wang and Q. Ji, “A unified probabilistic approach modeling relationships between attributes and objects,” in *IEEE International Conference on Computer Vision*, 2013. [33](#)
- [175] A. Kovashka and K. Grauman, “Attribute adaptation for personalized image search,” in *IEEE International Conference on Computer Vision*, 2013. [33](#)
- [176] W. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, “Multi-attribute spaces: Calibration for attribute fusion and similarity search,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [33](#), [90](#), [91](#), [115](#)

REFERENCES

- [177] O. Russakovsky and L. Fei-Fei, “Attribute learning in large-scale datasets,” in *European Conference on Computer Vision*, 2010. [34](#), [84](#), [85](#), [87](#), [90](#)
- [178] M. Sadeghi and A. Farhadi, “Recognition using visual phrases,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [34](#)
- [179] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [35](#), [36](#), [43](#), [46](#), [76](#), [85](#)
- [180] J. Philbin, J. Sivic, and A. Zisserman, “Geometric latent dirichlet allocation on a matching graph for large-scale image datasets,” *International Journal of Computer Vision*, vol. 95, no. 2, pp. 138–153, 2011. [35](#), [43](#)
- [181] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. [35](#)
- [182] L. Cao and L. Fei-Fei, “Spatially coherent latent topic model for concurrent object segmentation and classification,” in *IEEE International Conference on Computer Vision*, 2007. [35](#)
- [183] L.-J. Li, R. Socher, and L. Fei-Fei, “Towards total scene understanding: classification, annotation and segmentation in an automatic framework,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [35](#), [36](#), [55](#)
- [184] C. Wang, D. Blei, and L. Fei-Fei, “Simultaneous image classification and annotation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [35](#), [55](#)
- [185] N. Rasiwasia and N. Vasconcelos, “Latent dirichlet allocation models for image classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2665–2679, 2013. [35](#), [85](#), [124](#)
- [186] D. M. Blei and M. I. Jordan, “Modeling annotated data,” in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003. [35](#), [70](#), [123](#)

REFERENCES

- [187] R. Socher and L. Fei-Fei, “Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [35](#)
- [188] Z. Shi, T. M. Hospedales, and T. Xiang, “Bayesian joint topic modelling for weakly supervised object localisation,” in *IEEE International Conference on Computer Vision*, 2013. [35](#), [85](#), [87](#), [88](#)
- [189] F. Doshi-Velez, K. T. Miller, J. V. Gael, and Y.-W. Teh, “Variational inference for the indian buffet process,” University of Cambridge, Tech. Rep., 2009. [35](#), [76](#), [79](#), [80](#), [81](#)
- [190] J. Verbeek and B. Triggs, “Region classification with markov field aspect models,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. [36](#), [78](#), [81](#), [105](#)
- [191] B. Zhao, L. Fei-Fei, and E. P. Xing, “Image segmentation with topic random field,” in *European Conference on Computer Vision*, 2010. [36](#), [78](#), [81](#)
- [192] B. Zhou, X. Wang, and X. Tang, “Random field topic model for semantic region analysis in crowded scenes from tracklets,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [36](#), [78](#), [106](#)
- [193] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, “Matrix completion for multi-label image classification,” in *Advances in Neural Information Processing Systems*, 2011. [36](#), [56](#), [57](#)
- [194] Z. Zhou and M. Zhang, “Multi-instance multilabel learning with application to scene classification,” in *Advances in Neural Information Processing Systems*, 2007. [36](#), [123](#)
- [195] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, “Joint multi-label multi-instance learning for image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. [36](#), [123](#)
- [196] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, “Describing visual scenes using transformed objects and parts,” *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 291–330, 2008. [36](#), [119](#)

REFERENCES

- [197] D. Kuettel and V. Ferrari, “Figure-ground segmentation by transferring window masks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [37](#)
- [198] A. Zweig and D. Weinshall, “Exploiting Object Hierarchy: Combining Models from Different Category Levels,” in *IEEE International Conference on Computer Vision*, 2007. [37](#)
- [199] M. Rohrbach, M. Stark, and B. Schiele, “Evaluating knowledge transfer and zero-shot learning in a large-scale setting,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. [37](#)
- [200] T. Pedersen, S. Patwardhan, and J. Michelizzi, “Wordnet::similarity: measuring the relatedness of concepts,” in *Conference of the North American Chapter of the Association for Computational Linguistics ?Human Language Technologies*, 2004. [37](#), [52](#), [58](#)
- [201] L. Cao, Z. Liu, and T. S. Huang, “Cross-dataset action detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. [37](#), [38](#)
- [202] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. [37](#), [38](#)
- [203] L. T. Alessandro Bergamo, “Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach,” in *Advances in Neural Information Processing Systems*, 2010. [37](#), [53](#)
- [204] L. Jie, T. Tommasi, and B. Caputo, “Multiclass transfer learning from unconstrained priors,” in *IEEE International Conference on Computer Vision*, 2011. [37](#), [38](#)
- [205] J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, “Semi-supervised domain adaptation with instance constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [37](#)
- [206] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, “Transferring naive bayes classifiers for text classification,” in *Association for the Advancement of Artificial Intelligence*, 2007. [38](#), [53](#)

REFERENCES

- [207] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *IEEE International Conference on Computer Vision*, 2013. [38](#), [110](#), [113](#), [115](#)
- [208] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in *European Conference on Computer Vision*, 2010. [38](#)
- [209] W. Bian, D. Tao, and Y. Rui, “Cross-domain human action recognition,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 2, pp. 298–307, 2012. [38](#)
- [210] B. Gong, Y. Shi, F. Sha, and K. Grauman, “Geodesic flow kernel for unsupervised domain adaptation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [38](#)
- [211] A. McCallum and K. Nigam, “Employing em and pool-based active learning for text classification,” in *International Conference on Machine Learning*, 1998. [38](#)
- [212] M. B. Blaschko, A. Vedaldi, and A. Zisserman, “Simultaneous object detection and ranking with weak supervision,” in *Advances in Neural Information Processing Systems*, 2010. [38](#)
- [213] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, “Discriminative segment annotation in weakly labeled video,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. [38](#), [64](#)
- [214] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar, “Weakly supervised learning of object segmentations from web-scale video,” in *European Conference on Computer Vision*, 2012. [38](#)
- [215] P. V. Gehler and S. Nowozin, “On feature combination for multiclass object classification,” in *IEEE International Conference on Computer Vision*, 2009. [38](#)
- [216] P. Gehler and S. Nowozin, “Let the kernel figure it out; principled learning of pre-processing for kernel classifiers,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [38](#), [39](#)

REFERENCES

- [217] F. Orabona, L. Jie, and B. Caputo, “Online-batch strongly convex multi kernel learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 38, 39
- [218] P. Siva, C. Russell, T. Xiang, and L. Agapito, “Looking beyond the image: Un-supervised learning for object saliency and detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 39, 56, 57
- [219] J. Winn and C. M. Bishop, “Variational message passing,” *Journal of Machine Learning Research*, vol. 6, no. 3, pp. 661–694, 2005. 48
- [220] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006. 48, 78
- [221] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, “Hough forests for object detection, tracking, and action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011. 50
- [222] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002. 55
- [223] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, “Unsupervised object class discovery via saliency-guided multiple class learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 56, 57
- [224] R. Cinbis, J. Verbeek, and C. Schmid, “Multi-fold ml training for weakly supervised object localization,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 57, 58
- [225] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 58
- [226] A. Vezhnevets and V. Ferrari, “Associative embeddings for large-scale knowledge transfer with self-assessment,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 63

REFERENCES

- [227] A. Papazoglou and V. Ferrari, “Fast object segmentation in unconstrained video,” in *IEEE International Conference on Computer Vision*, 2013. [64](#), [65](#)
- [228] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010. [75](#)
- [229] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011. [75](#)
- [230] G. Patterson and J. Hays, “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [84](#)
- [231] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg, “From large scale image categorization to entry-level categories,” in *IEEE International Conference on Computer Vision*, 2013. [85](#)
- [232] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008. [85](#)
- [233] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, “Multi-instance multi-label learning,” *Artificial Intelligence*, vol. 176, no. 1, pp. 2291 – 2320, 2012. [85](#), [87](#)
- [234] Z. Feng, R. Jin, and A. Jain, “Large-scale image annotation by efficient and robust kernel metric learning,” in *IEEE International Conference on Computer Vision*, 2013. [87](#)
- [235] L. Wu, R. Jin, and A. K. Jain, “Tag completion for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 716–727, 2013. [87](#)
- [236] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, “Deformable part descriptors for fine-grained recognition and attribute prediction,” in *IEEE International Conference on Computer Vision*, 2013. [89](#)

REFERENCES

- [237] J. Dong, Q. Chen, S. Yan, and A. Yuille, “Towards unified object detection and semantic segmentation,” in *European Conference on Computer Vision*, 2014. [92](#)
- [238] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, “Unsupervised object discovery: A comparison,” *International Journal of Computer Vision*, 2009. [92](#)
- [239] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using amazon’s mechanical turk,” in *Workshop of the North American Chapter of the Association for Computational Linguistics ?Human Language Technologies*, 2010. [93](#)
- [240] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” in *Advances in Neural Information Processing Systems*, 2011. [93](#)
- [241] J. Tighe and S. Lazebnik, “Superparsing,” *International Journal of Computer Vision*, vol. 101, no. 2, pp. 329–349, 2013. [95](#), [96](#), [97](#)
- [242] S. Gould, J. Zhao, X. He, and Y. Zhang, “Superpixel graph label transfer with learned distance metric,” in *European Conference on Computer Vision*, 2014. [95](#), [96](#)
- [243] A. Vezhnevets, V. Ferrari, and J. Buhmann, “Weakly supervised semantic segmentation with a multi-image model,” in *IEEE International Conference on Computer Vision*, 2011. [95](#)
- [244] I. Olonetsky and S. Avidan, “Treecann - k-d tree coherence approximate nearest neighbor algorithm,” in *European Conference on Computer Vision*, 2012. [107](#)
- [245] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007. [108](#)
- [246] W. Li, R. Zhao, and X. Wang, “Human reidentification with transferred metric learning,” in *Asian Conference on Computer Vision*, 2012. [108](#)

REFERENCES

- [247] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, “Mahalanobis distance learning for person re-identification,” in *Person Re-Identification*. Springer, 2014. [108](#)
- [248] Y. Deng, P. Luo, C. C. Loy, and X. Tang, “Pedestrian attribute recognition at far distance,” in *ACM International Conference on Multimedia*, 2014. [109](#), [114](#), [115](#), [131](#)
- [249] Y. Yang and D. Ramanan, “Articulated human detection with flexible mixtures of parts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013. [109](#)
- [250] Q. Zhang, X. Shen, L. Xu, and J. Jia, “Rolling guidance filter,” in *European Conference on Computer Vision*, 2014. [109](#)
- [251] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *IEEE International Conference on Computer Vision*, 2003. [114](#)
- [252] A. Babenko, A. Slesarev, A. Chigorin, and Lempitsky, “Neural codes for image retrieval,” in *European Conference on Computer Vision*, 2014. [114](#)
- [253] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu, “Correlative multi-label multi-instance image annotation,” in *IEEE International Conference on Computer Vision*, 2011. [123](#)